

Collaborations essential for food in the developing world

The stakes in the debate on genetically modified crops increase dramatically as it moves to the developing countries. Governments need to help existing agencies to rise to the challenges.

To stand idly by while 40,000 children a day are dying from diseases related to malnutrition would, as was rightly pointed out at a meeting on agricultural biotechnology in Washington last week, be scandalous behaviour. But to agree with that sentiment is not to infer that the immediate, large-scale commercialization of transgenic crops is necessarily in the best interests of the world's poor.

The meeting, which was organized by the influential Consultative Group on International Agricultural Research (CGIAR), was called "Ensuring food security, protecting the environment, reducing poverty in developing countries: Can biotechnology help?", and the answer to the question is obviously 'yes' (see page 831). Transgenic technology clearly has the potential, for example, to enhance the nutritional value of rice in ways that might help balance the diets of hundreds of millions of people who depend on the crop.

But there are many approaches that plant scientists can take to improve subsistence crops, and that politicians can take to strengthen subsistence agriculture, which do not require transgenic technology. The earliest manifestations of the technology — crops that produce their own insecticides or resist expensive weed-killers — have little relevance to farmers in most of the developing world. The immediate priority for these countries is to obtain the best possible information about this rapidly evolving field, so that their governments can make informed choices endorsed by public support. That is where guidance from organizations such as CGIAR will be essential. The group, which coordinates the activities of 16 major food-research laboratories in developing countries, has the credibility needed to help Third World countries make adequately informed decisions.

These laboratories led the first Green Revolution, which transformed subsistence agriculture by giving poor farmers access to better seed and basic technology, and doubled the global cereal crop between 1960 and 1990. But that revolution was pioneered by plant scientists working in publicly funded laboratories. The next

transformation of agriculture is being driven by industrial corporations. It is difficult to see how the interests of poor farmers will necessarily be protected during this transformation, particularly as most of the research and development currently under way is aimed at intensive commodity farming, primarily in the industrialized world. Furthermore, whereas the Green Revolution was based on freely available breeding techniques, transgenics technology involves a spider's web of patent-protected information.

Unless the corporations holding most of that information can win the trust of the developing countries, their revolution will fail. Some business leaders, to their credit, have been calling for a stronger public-sector research effort in agricultural biotechnology, and philanthropic organizations such as the Rockefeller Foundation are providing important support. But, given the lack of available public resources, it may fall to the corporations to help pay for such an effort.

Nevertheless, governments in the industrialized world must acknowledge that organizations such as CGIAR can help developing countries to manage transgenic technology, and resource these groups accordingly. The standard policy at the moment (best exemplified by the United States) is to curtail foreign aid, warning that it creates 'dependency', while lavishing vast subsidies on domestic farmers, who in turn crush the more advanced commodity farmers of the developing world beneath a mountain of subsidized grain.

Governments in the developing world face an immense challenge in deciding what to do with the new technology. They can start by avoiding the mistake the industrialized world made by commercializing genetically modified crops without a genuine attempt at first winning public trust. They will need to work more closely together than they have in the past, through CGIAR and other forums. Ultimately, they may take the lead in applying a technology that, despite its early false steps, should play a significant part in alleviating world hunger. ■

Code for fairer lab management

A debate on the prospects for women in science reveals some pessimism as well as constructive ideas of wider relevance.

Sexual discrimination was a precursor of the spy scandal at Los Alamos (Cheryl Fillekes). Children are more interesting than fleas (Miriam Rothschild). These are among the most striking statements that emerged during *Nature's* six-week online discussion of the problems facing women in science, which ended last week (see <http://helix.nature.com/debates/women>). (To do her justice, Rothschild was making the more serious point that significant science can be done at home.)

But there was much else besides: how being a mother can lead to enhanced research productivity; the policy of telecommunications giant Ericsson of encouraging men to spend time at home with their children on the principle that this makes them better managers; a woman in a pharmaceutical company describing the "grooming of the

male establishment-in-waiting", thanks to chauvinism and cronyism; and disagreement about men being recruited to promote women's interests. Amid complaints that men are unwilling to share family burdens, there were examples of constructive initiatives within particular institutions and at a national level in several countries.

One productive way forward (Samuel Gorovitz) would be to build on codes of practice such as the National Academy of Sciences' "On being a scientist". A better code would also cover a range of issues that arise in laboratories and universities — including hierarchy, mentoring and the freedom to question, as well as sexual discrimination. Such a code, effectively backed, might leave no hiding place for those who discriminate unjustly. It should be high on the agenda of any laboratory that lacks it. ■





Malaria alert
Japanese firms join
the campaign against
a resurgent disease
p832



Mouse work
A popular laboratory
animal gets a
phenotype database
p833



PC feelings
MIT symposium
anticipates the
sensitive computer
p836



New broom
Profile of Europe's
new research
commissioner
p837

Developing countries look for guidance in GM crops debate ...

Washington

Developing countries have called on the Consultative Group on International Agricultural Research (CGIAR) — an influential federation of agricultural research centres — to develop guidelines for the research, trial and commercialization of genetically modified (GM) crops.

At a meeting in Washington last week, hosted by CGIAR and the US National Academy of Sciences, agricultural researchers and research administrators appealed to CGIAR to provide guidance to help poorer countries address the global debate over the application of agricultural biotechnology.

With consumers' groups, largely from Europe, and purveyors of transgenic crops, mainly in the United States, battling to determine the global acceptability of the new technology, many speakers warned that the interests of poor nations are being brushed aside.

Even the most powerful developing countries are seeking help from CGIAR, a network of 16 major agricultural centres sponsored by the World Bank, the Food and Agriculture Organization and the United Nations, which spent \$340 million last year on agricultural research.

Manju Sharma, secretary for biotechnology at India's Ministry of Science and Technology, called on CGIAR to publish guidelines on scientific research, field trials and commercialization to help governments set policies on agricultural biotechnology.

Speakers at the meeting also said that developing countries will depend on CGIAR to help counter the influence of the private corporations that control patents and information on transgenic crops.

Viloo Morawala-Patell, a professor at the University of Agricultural Sciences, Bangalore, India, said that public resistance is "not so much to GM food as to big industry". She called on CGIAR to "set up a parallel and alternative technology base" to that established by the corporations "in which the status of the farmer is protected".

CGIAR demonstrated its influence on the global GM food debate earlier this year, when it called on developing countries to boycott



Pressure point: protestors in Colombia lobby talks on regulating GM trade earlier this year.

the 'Terminator' gene technology, which Monsanto has since abandoned.

CGIAR held the meeting of several hundred of its centres' officials and other inter-

ested parties to help develop its approach to transgenic crops. "Everyone is waiting for them, because they are such a big actor," says Calestous Juma, a professor at Harvard University and a special adviser to the group.

"They can't sit on the fence anymore, because everyone is hacking away at the fence." But the "conflicting interests" of CGIAR's donors, who include the major industrialized countries, make it hard to determine a policy on GM crops, says Juma.

Older biotechnology tools, such as genetic markers in plant breeding, are firmly established in the centres. But less than ten per cent of their work currently involves transgenics, say officials.

Many of the research centres are determining whether and how to take their first transgenic crops into field trials. As the Washington meeting took place, for example, rice farmers were invited to the International

... as Rockefeller head warns of backlash

San Francisco

Public opposition to agricultural biotechnology in the industrial world could rob developing countries of the fruits of genetic research that are vital to their survival, according to Gordon Conway, the president of the Rockefeller Foundation.

Conway, who came to the foundation 18 months ago from a position as vice-chancellor of the University of Sussex in Brighton, England, is a renowned agricultural ecologist. He urges biotech multinationals to do more to address the ethical, economic, environmental and safety issues posed by crop manipulation, before a hostile public shuts down their operations.

He argues that the genetic engineering of food poses risks — such as outcrossing into wild

species and the creation of new viruses — that should be examined more closely. But he also emphasizes the potential benefits.

In an effort to address these issues, the Rockefeller is committing more than \$1 million a year to fund projects that foster constructive dialogue. In particular, the New York-based foundation aims to help developing countries become better informed and take a stronger role in policy discussions, so they can decide for themselves what level of risk is appropriate without becoming guinea-pigs for wealthier nations.

For example, the foundation recently gave \$260,000 to the African Centre for Technology Studies in Nairobi for a two-year project intended to help six

African governments develop their positions on genetic engineering and biodiversity.

Conway says the biotechnology industry has shown a new willingness to respond to argument. Monsanto, for example, recently promised not to develop the 'Terminator' technology that would make its seeds sterile and force farmers to buy new ones every year.

In early October, Monsanto's chief executive Robert Shapiro acknowledged the need to find common ground with biotechnology's opponents. He told the Greenpeace Business Conference in London that, until now, Monsanto had "irritated and antagonized more people than we have persuaded".

Du Pont, based in Wilmington, Delaware, plans to form an

► Centre for Tropical Agriculture (CIAT) at Cali, Colombia, to discuss trials of the centre's first transgenic rice plant, resistant to the hoja blanca virus that damages rice crops in Latin America.

Such trials are strongly supported by researchers and research administrators across the CGIAR network. But private corporations have found them difficult to put into practice. In Brazil an injunction by Greenpeace has stalled Monsanto's plans to test five breeds of soybean, and in Mexico concern has focused on the effect of GM maize on wild strains of the plant.

The extent to which the publicly funded CGIAR network should support either field trials or the commercialization of GM crops was fiercely debated. Brian Johnson of English Nature called for a moratorium on commercialization, and Fred Gould, an ecologist at North Carolina State University, warned that developing countries are ill-equipped to cope with unforeseen environmental problems that may arise from the crops.

But supporters of transgenic technology, such as Klaus Leisinger of Novartis, accused detractors of delaying nutritional improvements that could save thousands of lives.

Leisinger attacked what he termed "bio-McCarthyism". But Mark Sagoff, an ethicist at the University of Maryland, accused Leisinger of "fundamentalism" and argued that poverty is the real cause of malnutrition.

There was agreement with Sagoff's point that the 'safety' of GM crops is not the primary issue. As James Cook, a plant pathologist at Washington State University who represented the US National Academy at the meet-

ing, put it: "This whole debate isn't really about safety. Safety is the card which is played to get the deeper political and economic issues on to the table."

These issues include the fact that none of the first-generation transgenic crops are of much use to farmers in poor countries — rather, they will extend the productivity advantages enjoyed by heavily subsidized farmers in industrialized countries.

Another issue is the lack of technical knowledge in poor countries. But the most pressing concern is the imbalance of negotiating strength between the corporations that pioneered transgenic crops and farmers, scientists and governments in poor countries.

The developing world "must rely on the international organizations" to protect its rights, says Behzad Ghareyazie, director of the Agricultural Biotechnology Research Institute of Iran. But some speakers doubted whether even CGIAR has much negotiating power compared with the corporations.

Richard Jefferson, director of the Centre for the Application of Molecular Biology in International Agriculture in Australia, called on CGIAR to give its researchers "freedom to operate" in the face of ever-tightening restraints on their work.

Ismail Serageldin, chairman of CGIAR, was absent — he was in Paris, seeking to become head of the United Nations Education, Scientific and Cultural Organization (see page 833). Alex McCalla, director of rural development at the World Bank, summed up on his behalf, saying: "We've heard nothing that shakes my conviction that biotechnology has tremendous potential." **Colin Macilwain**

Japanese companies join international fight against malaria

Tokyo

Twelve Japanese pharmaceutical companies announced this week that they are joining with the World Health Organization (WHO) and Japan's Ministry of Health and Welfare to launch an initiative to identify potential drugs against malaria.

JPMW — which stands for Japanese Pharma, Ministry of Health and Welfare and WHO — will become part of 'Roll Back Malaria', a programme launched last year by WHO to create a global strategy for controlling malaria.

JPMW's activities, to be jointly funded by WHO and Japan's health ministry, will complement other initiatives in the Roll Back Malaria programme. Among these is the New Medicines for Malaria Venture, a public/private sector project supported by funding agencies and drugs companies including Glaxo Wellcome and Hoffman-La Roche (see *Nature* 395, 417; 1998).

Also active in this field is Multilateral Initiatives on Malaria, a consortium that includes the US National Institutes of Health, the Wellcome Trust, the World Bank and other UN agencies, which aims to develop malaria research in Africa.

The recent increase in collaborative ventures is a response to the fact that malaria parasites are becoming increasingly resistant to existing antimalarials, while the pharmaceutical industry has virtually abandoned research on tropical diseases (see *Nature* 386, 540; 1997).

The companies involved in JPMW — Takeda, Eisai, Yamanouchi, Chugai, Shionogi, Fujisawa, Sankyo, Daiichi, Sun- tory, Yoshitomi, Dainippon and Sumitomo Pharmaceuticals — will provide compounds from their chemical libraries for antimalarial screening.

The molecules will be screened by the Kitasato Institute in Tokyo, which plans to carry out random testing of more than 12,000 molecules over the next five years. Kitasato will also screen more than 2,000 molecules from its own libraries. Any molecule showing activity in the screening will be followed up by WHO's Programme on Tropical Diseases Research.

Win Gutteridge, the programme's chief of product research and development, says that the tools are already available to reduce the burden of malaria mortality and morbidity. But halving the figures within a decade, the declared aim of the Roll Back Malaria programme, "will only be possible if new tools, including anti-malarials, are developed". **Asako Saegusa**

► advisory group to review its activities in agricultural biotechnology and to provide a critique of the company's performance.

Conway praises these moves as "a good first step". But he urges the companies to do more, for example by donating enabling technologies to developing countries, by accepting 'plant variety' protection instead of seeking patents, and by using part of their profits to assist public-sector research.

Labelling should be considered a 'freedom of information' issue, says Conway, in that the public has a right to know what it is eating and to choose whether to buy genetically engineered foods. Companies may decide not to develop certain technologies because of social concerns, he adds. He urges governments, corporations, activists and scientists from the



Conway: fears public hostility to GM will harm development.

developed and the developing world to pinpoint significant issues and negotiate solutions.

Behind Conway's suggestions is a concern that public opposition to private initiatives could undermine public programmes, such as those financed by the Rockefeller Foundation.

"We are concerned that the public sector science is being constrained on the one end by limited access to the technologies,

and on the other by increasing wariness about getting involved because [biotechnology] is so unpopular," said Gary Toennissen, deputy director of agricultural services for Rockefeller.

Crop research that could help poor countries is not being pursued by companies, and some of the aid donors are becoming increasingly squeamish about it, he says.

The Rockefeller Foundation has spent more than \$100 million on plant biotechnology research, focusing on helping the poor. Scientists receiving its funds recently announced that they had used genetic modification to make rice produce β -carotene (which is converted to vitamin A) and iron, nutrients lacking in the diets of developing countries. After checks on its environmental and human effects, the rice will be donated to developing countries. **Sally Lehrman**

Database to standardize mouse phenotyping

Munich

A pilot study to systematically characterize 21 commonly used inbred — and thus genetically homogeneous — strains of mice has been launched by scientists in several US laboratories. The study is the first action of a project called the Mouse Phenotyping Initiative, which aims to create a publicly accessible database of the characteristics — known as traits or phenotypes — of inbred mouse strains.

According to members of the initiative's steering group, chaired by Ken Paigen, director of the Jackson Laboratory in Bar Harbor, Maine, such a database is urgently needed by mouse geneticists, as there is currently a lack of information about the strains that are routinely used to model human diseases.

Paigen is the driving force behind the initiative, which will set up the database at the Jackson Laboratory, the world's leading supplier of inbred mouse strains. The laboratory already hosts the Mouse Genome Database, which is accessible to researchers throughout the world.

With the Human Genome Project nearing completion, scientists are becoming increasingly interested in investigating the functions of the genes that have been sequenced, and mice have become a classic genetic tool in this endeavour.

Scientists around the world are due to finish sequencing the mouse genome by 2005. The speed of its completion was boosted by the creation earlier this month of the Mouse Genome Sequencing Network, funded by the US National Institutes of Health (NIH). The network involves ten collaborating laboratories which will share US\$21 million over the next six months.

In addition, there are already hundreds of relevant inbred strains and thousands of mutant strains, created either by 'knockout' technologies, which target particular genes for deletion, or by chemical mutagenesis.

Geneticists use inbred strains to analyse complex physiological or behavioural traits, such as a propensity to a particular type of cancer, or a taste for alcohol. They can start to pin down the genes that determine the trait by crossing a strain that strongly displays a particular trait with one that displays it weakly, or not at all, and working back from genomic information. A more precise understanding of the phenotypes of inbred strains will allow experimenters to select the most appropriate strains for cross-breeding.

The Mouse Phenotype Initiative's steering group, which has two European, one Japanese and six US members, has drawn up a list of inbred strains and phenotypes that it thinks should be included in the database. It recommends that 20 animals from each strain be analysed in two different laborato-



Hearing aid: a mouse is tested for hereditary deafness, giving clues to its genetic origins.

ries with expertise in a particular area, such as behaviour or clinical chemistry.

The steering group has suggested that data already published should not be included, because strains bred in different laboratories sometimes undergo 'genetic drift' and so can no longer be assumed to be identical to the reference strain from which they were derived. The Jackson Laboratory will provide all the mice that will be phenotyped for the database.

The pilot study, which will be conducted at the Howard Hughes Medical Institute at Northwestern University in Chicago, Baylor College of Medicine, and the Jackson Laboratory, will study three characteristics: circadian rhythms, haematology and the response to a high-fat, high-cholesterol diet.

Funding for the pilot study has come from donations from pharmaceutical companies. Rick Woychik, director of the Parke-Davis Laboratory for Molecular Genetics at Oakridge, California, one of the donors, says the database will be an "immensely powerful

tool" for industry, which needs optimal mouse models to identify genes involved in human diseases. But long-term financing has not yet been secured.

Major funding agencies, including the NIH, Britain's Medical Research Council and France's INSERM, will be asked by the initiative's steering committee to fund laboratories involved in phenotype assessment and database management. Pharmaceutical companies will also be approached.

If such funding is secured, later phases of the project will include up to 50 different strains of mice and all phenotypes for which there is interest in the research community.

"The Mouse Phenotype Initiative is important because we need to develop comprehensive, common and systematic phenotyping tests," says Steve Brown, director of the Mammalian Genetics Unit of Britain's Medical Research Council. "In particular, clinical vocabulary should be standardized to the mouse, since the mouse is used to model human disease."

But mouse pathologists are rare, he points out. This was recognized last year by the NIH, which is set to fund several mutagenesis centres. These will focus on recessive genetic traits, to complement the European focus on dominant traits, and will be grouped into two areas: behaviour and development.

Maja Bucan, a geneticist at the Center for Neurobiology and Behavior at the University of Pennsylvania, who runs a mouse mutagenesis programme screening for behavioural traits, says "we all went into mutagenesis studies without information about how phenotyping should be standardized, and now we have to go back and do it. This is exactly why [the Mouse Phenotype Initiative] is so important."

Alison Abbott

Japan wins bid for top Unesco post

London

Koichiro Matsuura, Japan's ambassador to France, is expected to be officially appointed next month as the new director-general of the United Nations Educational, Scientific and Cultural Organization (Unesco).

This follows his nomination last week by the UN agency's executive board, after Matsuura obtained 34 of the 58 votes in the third round of voting, a wide lead over his two closest rivals, Ghazi Algosaibi of Saudi Arabia and Ismail Serageldin of Egypt, a vice-president of the World Bank.

In an interview with *Nature* earlier this year, Matsuura promised that a more efficient bureaucracy and greater efforts to persuade the United States to rejoin the



Matsuura: seeking to foster consensus.

organization would be high among his priorities (see *Nature* 398, 554; 1999). "US re-entry is very important," Matsuura said. "We cannot hope to have meaningful discussions without the United States as an active participant."

He also said that Japan was keen to bring its consensus-style approach to problem solving to international issues. He pointed, for example, to the conference of parties to the UN climate convention, held

two years ago in Kyoto, at which Japan, together with the European Union, successfully steered a middle course between the positions of the United States and developing countries.

Matsuura's appointment is the result of an intense campaign by the Japanese government, which is keen to raise its profile in international affairs. He had the active support of physicist Akito Arima, formerly the president of the University of Tokyo, and until recently his country's minister of education and research. And the Japanese prime minister, Keizo Obuchi, an old schoolfriend, is said to have lobbied several European heads of state on Matsuura's behalf.

This campaign has led to speculation that Japan may have promised additional aid to some developing countries in return for their support for Matsuura's nomination. Matsuura has extensive experience of dealing with such countries, having been an architect of Japan's aid policies during the 1980s.

But he is said to have performed only moderately well in his interview with the executive board, and his appointment is widely seen as a vote for the Japanese government, rather than the ambassador personally. Japan is Unesco's largest donor country, and contributes 25 per cent of the organization's \$544 million annual budget. **David Dickson and Peter Pockley**

US Congressman boycotts science meeting with China

San Diego

The first in a planned decade-long series of meetings between US and Chinese scientific leaders was marred this week by the withdrawal of a powerful US Congressman, citing allegations of Chinese spying and illegal technology uses.

James Sensenbrenner (Republican, Wisconsin), chairman of the House Science Committee, cancelled his plans to attend the Sino-US Joint Science Policy Seminar as the delegation was leaving for the four-day meeting with representatives of the National Natural Sciences Foundation of China, due to open in Beijing last Sunday (24 October).

Sensenbrenner contacted Rita Colwell, the director of the National Science Foundation (NSF), the meeting's sponsor, asking her to cancel the delegation's trip because of continued Chinese provocations. "A seminar with senior officials from the White House and Congress could convey to the Chinese a business-as-usual attitude in our relationship on science and technology," Sensenbrenner said in a statement.

The Congressman was angered by the indictment earlier this month of the China National Aero-Technology Import and Export Corporation and the McDonnell

Douglas Corporation for the alleged illegal transfer of US defence technology to a Chinese missile factory.

He also quoted allegations of Chinese spying involving nuclear bomb secrets from the Los Alamos National Laboratory in New Mexico (see *Nature* **399**, 395; 1999), and "mischaracterizations" by Chinese officials visiting Congress this month for the first time since the Tiananmen Square conflict.

But Colwell declined to call off the trip, saying in a statement that "I've consulted widely with colleagues within and without [President Bill Clinton's] administration, and there is unanimity that the seminar is not linked to Sensenbrenner's specific concerns".

Colwell said the event "reflects the principle of free circulation of scientists and our continuing commitment to an open discussion of issues of common scientific interest".

US participants at the seminar included representatives from the NSF, the National Institutes of Health, the National Academy of Engineering and several universities. The latter included Richard C. Atkinson, president of the University of California — ironically, the institution that manages the Los Alamos laboratory from which nuclear secrets are alleged to have been stolen. **Rex Dalton**

European biologists unite to lobby for more money

Munich

A group of leading European molecular biologists is to launch a new initiative, called the European Life Sciences Forum (ELSF), to present a united view to politicians of the needs of the basic research community in Europe.

Frank Gannon, director of the Heidelberg-based European Molecular Biology Organization (EMBO), told the Munich Symposium on Cell Dynamics last week that the forum was necessary because of the fragmentation of the community's voice.

This fragmentation had weakened the impact of researchers on the European Union's fifth five-year Framework programme of research (FP5), which started this year, he said. "EMBO was one of very many life sciences organizations in Europe which individually advised the European Commission on how it should structure the programme, yet we saw no evidence of EMBO's input in the final design."

Given the "exponentially increasing" economic and social importance of life sciences, and the increasing demands of the biotech-



Gannon: life sciences need a single voice.

nology industry for advances in basic research, the life sciences need "a step up in funding, not incremental increases, as is currently the case in Europe," said Gannon.

One of the first tasks of ELSF, whose manifesto will be published on the Internet, will be to establish dialogue with the European Commission and coordinate a single input into the sixth Framework programme on behalf of all European basic researchers in the life sciences when such discussions begin in the next two years.

Don Cleveland, a professor at the University of California at San Diego, told the Munich meeting that a recent commitment to doubling the budget of the National Institutes of Health by the US Congress had been largely due to a campaign by scientists in the US societies for cell biology, biochemistry and biophysics.

Ten years ago, said Cleveland, when the American Society of Cell Biology was still very small, it started to spend a significant proportion of its budget on paying a full-time member of staff to "worry about public advocacy", as well as the services of a professional lobbyist in Washington. "At the time, some scientists said we were spending too much," he said. "But it paid off."

Other scientists at the meeting pointed to the success of Greenpeace's well-funded campaigns in Europe — such as that against genetically modified foods — which they said urgently needed to be matched by similarly competent campaigns by scientists.

Gannon pointed out that the task of lobbying is more complicated in Europe than in the United States. "We have to lobby in ten different languages and in twice as many political centres of power," he said.

But he said that scientists had themselves to blame for inadequate funding of life sciences in Europe "because we have never been able to come up with the simple sort of message that politicians can understand and convey to their constituents". **Alison Abbott**

Ukrainian scientists charged over transfer of data to West

London

The Ukrainian Security Service has arrested a prominent marine biologist, Sergei Piontkovski of the Institute of Biology of Southern Seas (IBSS) in Sebastopol, on charges of transferring secret information abroad and handling hard currency illegally. Four other scientists from the institute are facing arrest on the same charges.

Files and computers from the institute have been seized, and it is understood that the five scientists are being interrogated by the security service. Piontkovski, who recently returned from a period of research at Stony Brook in the United States, is due to appear in court in about a month, and could face up to 20 years in prison.

Piontkovski, an internationally known marine biologist, has received funding from international organizations and foreign governments. These include the European aid programme INTAS, the US Office of Naval Research — for a project on bioluminescence — and the British government, which paid for a plankton biodiversity project.

The British grant, funded through the Darwin Initiative, has come under particular scrutiny. The scientists are being questioned about what is being described as the “criminal transfer” of Ukrainian scientific information to the west, as well as the transfer to their institute of hard currency from Britain.

Piontkovski's work with institutions such as the Plymouth Marine Laboratory and the Royal Society in Britain, Amsterdam University in the Netherlands, and the Office of Naval Research and the Smithsonian Institution in the United States, is under scrutiny.

Robert Williams of the Plymouth Marine Laboratory, the principal scientist on the Darwin Initiative project, has expressed concern at the actions of the Ukraine authorities. He points out that all the data for the international projects was biological information collected in international waters between 1950 and 1989.

“It has no strategic importance whatsoever,” says Williams. Piontkovski appears to have been collecting plankton species data, temperature, salinity and density data.

Meanwhile the Ukrainian Security Service has begun searches in another scientific organization, the Institute of Marine Geophysics (MHI), situated some 200 metres from IBSS. Last Saturday (23 October) the Sebastopol scientists launched a campaign to collect signatures to ask the Ukrainian president Leonid Kuchma to intervene. It is understood that the directors of the IBSS and the MHI are taking up the issue with the Ukrainian National Academy of Sciences.



The Ukrainian embassies in Moscow and London declined to comment. But they confirm that the sentence for spying can be up to 15 years in prison. Illegal currency operations carry a five-year sentence, with confiscation of all belongings. The US embassy in Kiev was unable to confirm rumours on Tuesday that the US ambassador was to issue a statement on the situation.

“This will knock back science a decade, to pre-Soviet strictures,” says Williams. “People will not be prepared to collaborate.” Williams has written to NATO and to the Ukrainian academy warning that, if the situation cannot be resolved, conferences such as the European Marine Biology Symposium, scheduled for Sebastopol next autumn, could be cancelled.

Williams confirms that Piontkovski has consistently tried to direct funds to colleagues in the Ukraine looking for support and grants. He says that IBSS directors have encouraged staff — who are only being paid by the institute for two days work a week — to find external sources of funding.

Piontkovski is reported to have been receiving hard currency through his role as the project director in the Ukraine for the Darwin Initiative project. He was responsible for purchasing equipment and distributing funds to IBSS colleagues. To date he has received one cash transfer from the Bank of England.

Some suggest Piontkovski was reported to the security service by a staff member. This came when moves were made to promote Piontkovski to head of department at IBSS.

The IBSS deputy director, Yuri Tokarev, who is one of the researchers under investigation, has written to the chairman of the state science committee to defend himself and his colleagues. “I hope that common sense will win, and that our contacts with the foreign scientists will not be broken by the security officers,” he wrote.

Natasha Loder & Carl Levitin

Indian government faces controversy over new IT ministry

New Delhi

India's coalition government, led by the Bharatiya Janata Party since this month's general election, has stirred controversy with plans to create a broad ministry for information technology (IT), to be headed by a non-technical person.

Because of the birth pains and confusion over its terms of reference, the ministry, whose creation was announced last week, will be under the control of the prime minister, Atal Behari Vajpayee, until a minister is appointed.

In the meantime, a senior official from the Indian Administrative Service has been appointed secretary to the ministry, causing concern in the country's IT industry.

The main obstacle to the creation of the ministry is likely to be existing ministries, whose authority and powers will be diluted as some of their responsibilities are transferred.

Government officials say that the IT ministry will promote Internet, e-commerce and knowledge-based industries without encroaching on existing ministries. But the Department of Electronics will come under the IT ministry, which will also take over the Internet and e-commerce services of the Ministry of Telecommunications.

Some of the functions of the Ministry of Information and Broadcasting — such as Internet-on-cable and cable television — will also be taken over. So will the National Informatics Centre (NIC), the country's first and largest government computer network.

N. Seshagiri, the founding director of NIC (see *Nature* 366, 622; 1993) says that, although he welcomes the new ministry, “it should be run, at least initially, by an IT professional”.

N. Chandrasekaran, director of the Centre for Development of Advanced Computing in Bangalore, argues that the ministry “will facilitate India's smooth integration into the global marketplace”. But Swami Manohar, a professor of computer science at the Indian Institute of Science in Bangalore, doubts that a separate IT ministry is needed.

“Given the convergence of computing, communication and broadcasting, the partitioning of the individual domains could be contentious, and the IT ministry may find itself spending most of its energy wrangling with other ministries to sort out the turf,” says Manohar.

K. S. Jayaraman

MIT promises us the sensitive computer ...

Boston

Computers will probably be able to sense human emotions in the near future, revolutionizing the relationship between humans and machines, a symposium sponsored by the Massachusetts Institute of Technology's (MIT) Media Laboratory heard last week.

Charles Vest, president of MIT, told the audience of about 1,000 — which included representatives from Ford, General Motors, IBM, Microsoft and Xerox — that “we’ll see a melding of the biological and physical worlds. Our machines are about to recognize our faces and talk back to us.”

Nicholas Negroponte, director of the laboratory, said that — as with many Media Lab projects — work on this goal is still at an early stage. “Everybody here is urged to be at the edge,” he said, pursuing research that “may or may not work out.”

But Marvin Minsky, an artificial intelligence pioneer based at the Media Lab, maintained that computers are still primitive. “Computers lack common sense,” he said. “Although computers started about 50 years ago, as far as I’m concerned they’re still just starting.”

The biggest problem with computers, said *Wall Street Journal* columnist Walter Mossberg, who hosted the day-long event, is that “we’ve tried to make people conform to stupid ways of doing things. We need to make our computers human-literate, rather than making us computer-literate.”

Rosalind Picard, head of the Media Lab’s Affective Computing group, suggested that computer designers have “completely left emotion out of the equation”. She added: “We need a balance in computing, and emotion is one piece of the picture.”

Picard and her colleagues have developed ‘emotionally supportive agents’ to ease frustration among computer users. Once a computer senses that a user is becoming aggravated — using a camera and pattern-recognition technology to analyse facial expressions, by evaluating the sounds (or screams) a person makes, by gauging how hard a person squeezes the mouse, or by measuring skin conductivity, heart rate, pulse and other physiological variables — an autonomous software ‘agent’ can offer assistance, encouragement or comfort.

In one experiment, people using computers with an ‘empathic’ agent spent more time on their machines and were more satisfied than those without.

A computer developed at the Media Lab can now recognize certain emotions with about 80 per cent accuracy, once a ‘signal’ for sentiments such as anger, grief, hate, love and joy has been determined — based, in part, on actors’ depictions.

Picard admitted that this only works

for a limited number of emotions. “But I don’t think it will take us 50 years to recognize emotions across the board, even though it took that long for speech recognition technology to become ‘speaker independent’.”

A different use of this technology, on display at the symposium, is in new forms of artistic expression. In one demonstration, a dancer wore shoes fitted with sensors that converted his steps into music. In another, Media Lab graduate student Teresa Marrin demonstrated the ‘Conductor’s Jacket’ equipped with motion sensors that enabled her to adjust musical output through arm gestures.

Commenting on the two approaches, David Durlach, president of Boston-based TechnoFrolics, noted that “it is infinitely easier to elicit emotional responses in humans with machines than it is to get machines to respond appropriately to human emotions.

“I don’t have to understand everything about emotions to evoke feelings in people,” said Durlach, who creates interactive, computer-controlled sculptures. “I can tell from people’s expressions whether it works.” On the other hand, “getting a computer to understand emotions by explicitly coding all the cues is incredibly complex and, if done poorly, could be annoying or boring.”

Such an approach, he said, is probably doomed to failure. “But an adaptable system that learns about emotions over time — in much the same way a child does — could allow rich interactions between



Curious or confused? These spectacles, developed at Media Lab, can help monitor the emotions revealed by a furrowed brow.

humans and intelligent machines.”

Few at the meeting questioned the value of emotionally savvy computers. But a cautionary note was raised by Jamie Hsu of the General Motors Corporation, who asked whether overemphasis on human-machine interactions might have a negative effect on interactions between humans.

Computers are already changing how people behave, especially children, said Hsu. “I just want to make sure that when computers become sensible 30 to 50 years from now, humans also stay sensible.”

Steve Nadis

... and may look to Ireland for expansion

London

The Irish government and the Massachusetts Institute of Technology (MIT) are near to closing a deal that would create a \$200 million research and teaching centre for information technology in Dublin, the first outpost of the US institutes Media Lab.

Securing a major new European research centre would be a coup for the Irish government. According to a government spokesperson, it has been investing heavily in educational technology in schools and colleges over the past few years to bring Ireland from the “third division” to the premier division” in

information technology.

The spokesperson confirms that negotiations have taken place over the past few months between the Irish Department of Education and Science and the MIT Media Lab, and describes these as “a recognition of where this country is going”.

The country’s four universities are expected to be involved in the deal, although no decisions have been taken on how the interaction with them will work. Reports suggest that the new centre will focus on Internet applications, particularly e-commerce and technology in education, as well as providing research and business ‘incubation’. A number of MIT research staff would

be expected to move to the centre, which would house up to 300 students, mostly postgraduates.

It is believed that the government has been asked for a commitment of \$40 million towards the project. Its input will be balanced against its expectation that the centre will attract new industry to the country, particularly high-technology companies. Government officials are also said to be hoping that the centre will help ensure continued growth in the economy by meeting technology skills shortages. The Media Lab has declined to comment on the reports.

Natasha Loder

Bringing a community-based vision to the heart of Europe's research

The European Union's new research commissioner hopes that his plans to promote science will win over critics in the European Parliament.

Brussels

When Philippe Busquin was minister of the budget and energy for the Walloon Region in his native Belgium, he introduced a law requiring utility companies to provide all citizens with a minimum level of electricity, whatever their ability to pay.

"It was a relatively small measure in itself, but it showed — like the butterfly flapping its wings that can trigger a tornado — that by changing something a little you can change the whole system," says Busquin, the European Union's new research commissioner.

Like his controversial predecessor, Edith Cresson, Busquin has deep roots in left-wing politics. He owes his appointment, confirmed last month after close questioning of his political past by the European Parliament, to having been a prominent member of Belgium's socialist party for many years.

Also like Cresson, Busquin is keen that the commission's research budget, for which he is responsible — about 3 billion euros (US\$3.2 billion) a year — should focus on projects intended to raise the quality of life in Europe as well as its economic competitiveness.

But the similarities stop there. Unlike his predecessor, Busquin has a scientific background. He took a physics degree at the Free University of Brussels (where he studied under the Nobel laureate chemist Ilya Prigogine), and spent several years as an assistant lecturer at the university and as a physics teacher at a teacher-training college before entering full-time politics in 1977.

Catalysis, not co-ordination

Whereas Cresson appeared keen to increase the commission's central control over its budget, Busquin says he sees himself as a catalyst rather than a coordinator of the activities of European researchers.

His political vision is community-based. He points, for example, to his creation of a group made up of representatives from petrochemical companies, local authorities and local people that discusses the environmental impact of the companies' activities.

The group grew out of a research project that Busquin carried out as part of his postgraduate work in the mid-1970s on the impact of industry on local communities. "Twenty years later, the group is still meeting," he says with pride.

And while Cresson treated the media



Busquin: keen to promote networks and increase the contact between European researchers.

with disdain — refusing a series of requests for interviews from this journal, for example — Busquin talks freely about his ambitions for his term of office, even if his words already echo official thinking in Brussels.

His top priority, he says, is to create a European scientific *espace*, broadly translated as a European scientific community. "The added value we can bring to Europe's research efforts, 96 per cent of which are still financed by member states, is through our ability to catalyse such an *espace*."

The main tool at his disposal is the fifth five-year Framework programme (FP5), which was adopted last year and runs until 2003. Not surprisingly, he endorses the strategic priorities already embodied in this.

For example, Busquin speaks enthusiastically of promoting European science through the use of networks. "We need to address the needs of the research community, and in particular how to increase its mobility and the intra-European contact between researchers," he says.

Another priority is to give the commission's research programmes a more human face. "I think that the FP5 is a bit too oriented to competitiveness," he says. "We should also give people a vision of how science can serve the citizen. It is very important that we give

the public a positive vision of science and what it can do for them."

Few are likely to challenge this, or his insistence that problem-oriented research must be complemented by open-ended projects. "Science must not be completely instrumentalized," he says. "There should be some research is carried out for its own sake."

More closely watched will be his determination and ability to grasp the thornier issues facing the commission, such as how to improve the effectiveness of its Joint Research Centres — a long-standing target of criticism by some member states.

Busquin speaks highly of some of the work being carried out at the centre at Ispra in Italy, which he visited last month, for example in food safety or the analysis of Earth images from space. But he acknowledges that "in some cases you will have to change the mission of institutions; some tasks need to be redefined," adding that "one must not waste public funds".

Raising awareness

Another hot topic is the extent to which the European Commission should fund infrastructure facilities through the Framework programme, particularly in the life sciences. Here Busquin says he is a believer in "a variable geometry", adding that: "Not everything should be paid for by Brussels."

Busquin has already had a baptism of fire, having had both his involvement in Belgian politics and his lack of experience in managing large-scale projects closely questioned by the European Parliament's research committee (see *Nature* 401, 104; 1999).

In the end, the committee decided to neither endorse nor oppose his appointment. Busquin admits that the encounter was "not very agreeable", although he adds that, as a politician, he is used to taking criticism.

But he talks enthusiastically about the tasks ahead, in particular raising public awareness about science and encouraging school pupils to take an active interest. "We have to give them a taste for science" he says.

"We need to find ways of bringing young scientists together, and we must also give them role models." Busquin says that as a physics student his own role model, as a scientist with a social conscience, was the US physicist Robert Oppenheimer. He is clearly keen to be seen in the same light.

David Dickson

A priority is to give the commission's research programmes a more human face.

Former government science official moves to CNRS

Paris Jacqueline Godet, formerly director of life sciences at France's Ministry for National Education, Research and Technology, is to head the life-science department at France's Centre National de la Recherche Scientifique



Godet: to carry out Allègre's reforms?

(CNRS), it was announced last week.

Godet, who replaces Jacques Samarut, has held several posts at laboratories shared between CNRS and the University of Lyons, and in 1990 was named director of the joint Molecular and Cellular Genetics Centre.

The CNRS has been at the heart of government efforts to restructure French science over the past two years (see *Nature* 397, 463; 1999). In particular, the science minister, Claude Allègre, and a parliamentary commission have called for the CNRS to be restructured and to increase its ties with universities. Allègre also seeks to boost biological research, which is seen as lagging behind its competitors.

UK urges action over global warming threat to rainforest

London The Amazon rainforest could be devastated by global warming, according to research published by the UK government last week. Previous models have not predicted such a severe effect on the forest. The model shows that increased temperature and carbon dioxide will initially result in a net increase in biomass until 2080, because of the increased growth of forests in northern latitudes. Food shortages, flooding and the spread of insect-borne diseases are also predicted.

The British government has called for swift action to ratify the Kyoto protocol as signatories prepare to meet in Bonn this week to discuss its implementation. But according to the model, substantial forest loss would still occur even if emissions were stabilized at three times the pre-industrial level.

New Mexico telescope gets green light

San Diego Initial funding for a \$40 million observatory in New Mexico's Magdalena Mountains has been approved by the US Congress. The Magdalena Ridge Observatory, in the federally funded Langmuir research area in one of the Northern Hemisphere's darkest locations, will house three 2.4-metre

telescopes at 10,600 feet linked to scan the skies with optical interferometry.

The New Mexico Institute of Mining and Technology will lead a consortium of universities operating the observatory, which will also be used by the US Army to track missiles at the White Sands Missile Range. The Department of Defense provided the initial \$3.5 million for construction planning of the observatory, which is expected to become operational in about three years.

Mammoth find triggers plan to cross with elephants

Moscow A team of Russian, French, American, Japanese and Dutch scientists in the Siberian town of Khatanga hope to free a mammoth 3 metres tall from its covering of ice. The work will be done in a cave in the permafrost, to which the gigantic block of ice was delivered by helicopter from the Chelyuskin peninsular.

The mammoth, which died 23,000 years ago, was found 6 metres under the surface. Scientists think that the permafrost, which preserved the mammoth's tissues, also left its DNA undamaged. They plan to extract the DNA and place it into an elephant's embryo, which they will implant into a female elephant. The possibility of cloning mammoths is also being discussed. The work is funded by the Discovery TV channel.

US academy to study lie-detector tests

Washington The US Senate has asked the National Institutes of Health (NIH) to commission a study from the National Academy of Sciences on the efficacy of polygraph testing. The tests may be used by the Department of Energy on thousands of employees at its nuclear-weapons laboratories.

Scientists at the laboratories have expressed deep misgivings about the tests, which they believe to have no scientific basis. Senator Jeff Bingaman (Democrat, New Mexico), who shares their concern, says that the “psycho-physiological phenomena” on which polygraph tests are based fall within the NIH’s technical remit, and attached the request to NIH’s annual funding bill.

NIH in line for massive funding bonanza

Washington The National Institutes of Health (NIH) could receive a funding bonanza next year that is even larger than expected, as negotiators from the House of Representatives and the Senate have agreed an appropriations bill that would increase its budget by \$2.3 billion to \$17.9 billion.

The proposed increase — larger than both the \$2 billion approved by the Senate and the

\$1.3 billion proposed by House appropriators — emerged as supporters of the NIH sought to shelter it from a plan by Republicans to cut 1.4 per cent from the budget of every agency in the federal government. But as budget negotiations between the Congress and the Clinton administration enter their final phase, the across-the-board cut may fail to materialize, leaving the NIH with the full, \$2.3 billion increase.

French charity fraud figures head to prison

Paris Jacques Crozemarie, the founder and ex-president of the French Association for Cancer Research (ARC), was sentenced to four years in prison and fined FF1.5 million (US\$250,000) last week for his role in an embezzlement scandal that cost the charity FF300 million (see *Nature* 379, 103; 1996 and 399, 724; 1999).

Crozemarie will appeal against the ruling and remain free until its outcome is known. Michel Simon, the former chief executive officer of International Development, a public relations company contracted by ARC, was sentenced to three years in prison and also fined FF1.5 million. A total of 19 others involved in the affair received shorter prison terms.

Chinese prawn baculovirus genome fully sequenced

Tokyo US and Chinese researchers have finished sequencing the genome of the C-type baculovirus, or prawn white spot baculovirus (WSBV), a pathogen of prawns that cost China 10 billion yuan (US\$ 1.2 billion) in 1993 (see *Nature* 397, 465; 1999).

The project was the work of the National Centre for Biotechnology Development, the National Human Genome Research Centre, the Third Institute of Oceanography and Shanghai GenCore Biotechnologies. The results indicate that WSBV may belong to a different subfamily from originally thought, or even a new virus family. The 305-kilobase sequence will aid the development of strategies for the prevention and treatment of prawn diseases.

Correction

The illustration used with the article ‘US warms to carbon sequestration research’ (see *Nature* 401, 315; 1999) should have been referenced to ‘Direct experiments on the ocean disposal of fossil fuel CO₂’ by Peter G. Brewer, Gernot Friederich and Edward T. Peltzer of the Monterey Bay Aquarium Research Institute, and Franklin M. Orr of Stanford University (see *Science* 284, 943–945; 1999).

Accurate radiometers should measure the output of the Sun

Sir—We should make use of new technology to improve on the poor state of knowledge of the constancy of the Sun's output. In 1994, the National Research Council of the US National Academy of Sciences called for improved accuracy in measurements that monitor the Sun's output. In its report¹, the council stated that, between 1980 and 1986, an estimated 0.1% decrease in the output of the Sun as part of the 11-year cycle just cancelled the effect on the climate produced by anthropogenic greenhouse-gas emission in the same period. This highlights the accuracy required to separate long-term anthropogenic effects from natural ones.

In the past, accurate absolute radiometry, even on the ground, was difficult, and the spread of the results from different satellite radiometers over the 20 years that total solar irradiance has been monitored from space is nearly 0.5% (Fig. 1a). These differences are consistent with an uncertainty of about 0.4% associated with each radiometer². No useful measurements were obtained from the one set of radiometers recovered after flight (from the SOVA satellite). In trying to interpret these data, the experimenters took the only route possible and applied adjustments to the various data sets to normalize their absolute levels. Figure 1b shows the resulting self-consistent

data set, which appears to show that the underlying output of the Sun has remained stable to within 0.02% over the past 20 years. But how reliable is this conclusion?

It is well known in metrology that it is not possible to say by how much a measured quantity is drifting over a long period unless measurements are made at the beginning and end of the period relative to a standard that is demonstrably stable, that is, one directly linked to the fundamental constants of physics. From the data in Fig. 1, taking account of the uncertainties in the calibration of the radiometers and those inherent in the normalization procedures, we believe that an uncertainty of not less than 0.3% must be associated with the apparent null underlying variation in the output of the Sun shown in Fig. 1b. In view of the sensitivity of climate change to systematic variations in solar output, such an uncertainty is not good enough.

Although national metrology institutes now have absolute radiometers using new technology with accuracies below 0.01%, they have not been used for these important measurements. We call on the solar physics and Earth resources communities to collaborate with these institutes to draw up a long-term programme of absolute radiometry in space, using the best technology, with the aim of improving the accuracy of these crucial measurements by at least an order of magnitude. We owe this to future generations, especially to climatologists and those who will use the results of climate studies.

We must be able to rely on the results of

all measurements related to climate studies. They provide the essential basis for advice to governments in formulating policy, the financial and human consequences of which will be enormous. Such measurements must be made in SI units, which are firmly linked to the constants of physics.

T. J. Quinn*, C. Fröhlich†

*Bureau International des Poids et Mesures, Pavillon de Breteuil, 92312 Sèvres Cedex, France

†Physikalisch-Meteorologisches Observatorium, Davos World Radiation Centre, Dorfstrasse 33, CH-7260 Davos, Dorf, Switzerland

1. National Research Council *Solar Influences on Global Change* (National Academy Press, Washington DC, 1994).

2. Fröhlich, C. & Lean, J. *Geophys. Res. Lett.* **25**, 4377–4380 (1998).

Finding the complete bioinformaticist

Sir—The UK Biotechnology and Biological Sciences Research Council (BBSRC) does seem to share Peter Campbell's view that interdisciplinary research teams are a good way to drive bioinformatics forward (*Nature* **401**, 321; 1999). Indeed, over the past few years, it has supported just this kind of research through a joint initiative with the Engineering and Physical Sciences Research Council. The reviewing panel has typically assigned high weightings to projects that represent collaborations between biology and computer-science departments.

The conclusion reached by Campbell, and perhaps by the BBSRC—that it is difficult to find researchers adept in both computer science and biology—is puzzling. I would consider myself as just such a person: a scientific 'jack of all trades'. No doubt, to most biologists I appear to be a computer scientist, and to most computer scientists I appear to be a biologist. I believe that I know at least enough about both fields to understand some of the important problems in biology and which computational methods might be most appropriately applied to at least some of them.

But are people like me as hard to find as Campbell believes? Certainly, in other fields, such as mathematics, physics, chemistry or astronomy, there is no shortage of scientists with skills in their own subject and in computation. Of course, there is still a wealth of opportunity for collaboration between these fields and the computer sciences.

So, are 'all-in-one bioinformaticists' really in such short supply? I did a straw poll of the software packages that my colleagues and I use regularly. I found 20 bioinformatics packages to be particularly vital to our research projects, ranging from sequence

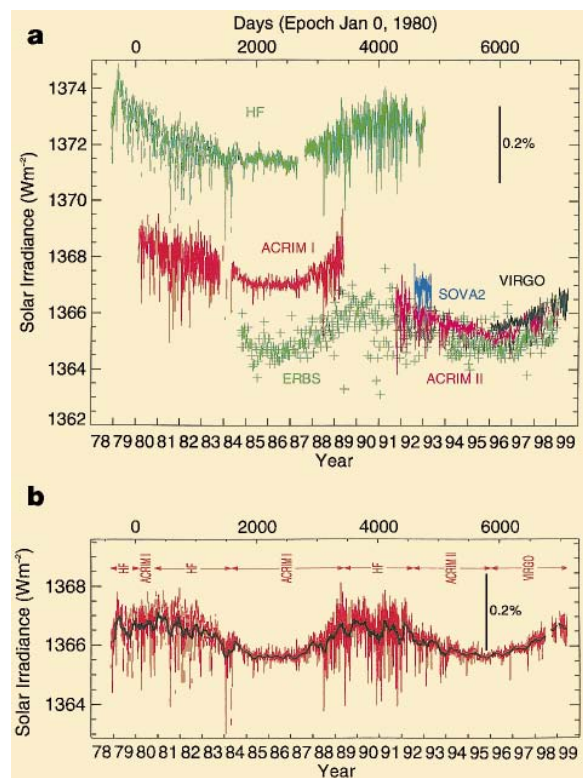


Figure 1a, Data from measurements of total solar irradiance made from satellites. b, Composite total solar irradiance obtained from the data shown in a by making adjustments to normalize the absolute levels of successive data sets. For details see ref. 2.

alignment programs to molecular graphics software. The principal or sole authors of 18 of them are computer-savvy biologists. Although this is an inconclusive experiment in statistical terms, the message is clear. Biologists with strong computer skills are certainly out there somewhere. So funding bodies should finance interdisciplinary research in bioinformatics, but must not forget the 'jacks of all trades' who have already made such a useful contribution.

David Jones

*Protein Bioinformatics Laboratory,
Department of Biological Sciences,
University of Warwick,
Coventry CV4 7AL, UK*

Pasadena pranks

Sir— I take exception to your unflattering characterization of the California Institute of Technology on the occasion of its being ranked the top US university by *US News & World Report* (Opinion, *Nature* **400**, 801; 1999). This ranking is not a fluke, even if caused by an arbitrary change in the criteria — the previous criteria were equally subjective.

The ranking signifies that Caltech, even though it is very small, has been, and will be, a power to reckon with. We have remained small by choice (only some 290 professors and 900 undergraduates) and do not aspire to the breadth of a larger university. But we do what we do extremely well, and manage to have a rich cultural life (and fun) while doing it.

That the magazine "had to delve back 15 years for an example of interesting non-curricular activity" is a failing of its research, not of the institute. A third of our students participate in intercollegiate sports, and student enterprises abound in music, theatre and the arts. Beyond such regular scientific visitors as Stephen Hawking, our campus has hosted recent visits by Tom Stoppard, Seamus Heaney, Walter Cronkite, Oliver Stone, Jonathan Miller, Beverly Sills and Warren Buffett.

The spectacular pranks that are part of our lore (such as changing the Hollywood sign to read "Caltech", or the Rose Bowl game prank you mentioned) stem not from football envy, but from the imagination and exuberance of our students, who request the 12 a.m.–2 a.m. recitation you mention to better manage their busy lives.

We are, as you note, listed as a poor "party school" because our students find fun in their own ways. And no one danced in the streets because we were too busy doing what we do best.

Jean-Paul Revel, A. B. Ruddock

*California Institute of Technology,
Pasadena,
California 91125, USA*

Hospital merger leaves clinical science intact

Sir— The News article by Rex Dalton about the merger between the hospitals of Stanford and the University of California, San Francisco (UCSF) is misleading (*Nature* **400**, 300; 1999).

First, it gives the impression that "faculty leaders" have abruptly changed course and are now calling for dissolution of the merger. Dalton quotes Warren Gold, who has opposed the merger from its outset. But many of our departmental chairs and other leaders of the faculty remain more open-minded about the fate of the merger. In particular, they recognize the substantial disadvantages now posed by dissolution, whatever their original views on the merger.

Second, Dalton appears to blame the merger for the pressures that increasingly impede clinicians from doing scholarly work. This is inaccurate. The pressures arise from the punishing realities of the medical marketplace. They existed before the formation of the merger, and they can be found at academic health centres throughout the United States.

Third, Dalton's article concludes with an undocumented assertion that basic scientists at UCSF are challenging the need for "clinical programmes". But no sensible basic scientist could imagine a medical school or health-science campus without clinical programmes.

I have been a member of the basic science faculty at UCSF for 30 years and know that it recognizes the importance of physician-scientists and clinical research. Indeed, the collegiality between basic scientists and clinicians at UCSF is exceptional.

J. Michael Bishop

*(Chancellor)
University of California, San Francisco,
513 Parnassus Avenue, S-126, San Francisco,
California 94143-0402, USA*

Sir— In the article entitled "Merger of top Californian medical schools turns sour", Dalton's suggestions that "the crisis is pitting physicians against basic science researchers" and that "Some basic scientists have even argued that clinical programmes aren't needed" are pure fantasy, as is the headline referring to the fictitious merger of the schools.

Merger of the schools has never been discussed, only merger of the hospitals, and we have not heard any of our basic science colleagues advocate the ludicrous notion of our medical school abandoning its clinical programmes.

Indeed, we basic scientists have been brought together with our clinical

colleagues in coping with the national crisis in funding for health care by institutions such as UCSF that are dedicated to the care of all patients, rich and poor, as well as to the creation of knowledge. Our current planning for the future of disease-related research reflects this fusion of interests and experience.

Michael P. Stryker*, Allan Basbaum†, Tony DeFranco‡, Ira Herskowitz§, Keith Yamamoto¶

*Departments of *Physiology, †Anatomy,
‡Microbiology and Immunology,
§Biochemistry and Biophysics
and ¶Cellular and Molecular Pharmacology,
University of California, San Francisco,
Parnassus Avenue, San Francisco,
California 94143-0444, USA*

Freedom to speak or to misinform?

Sir— There is no doubt that freedom of speech as well as freedom of science are values that need to be protected. But the News article about a dispute over freedom of speech at the Institute of Molecular Biotechnology (IMB) at Jena, Germany, seems to give credence to a small group of people whose main aim seems to be to discredit the institute (*Nature* **401**, 201; 1999).

The goal in the present case was not to limit Günter Löber's rights, but to prevent further dissemination of false allegations about his former place of work. It is important to ensure that the many current IMB employees retain the freedom to conduct their research. Löber was not prohibited from criticizing the IMB, but he has agreed not to repeat groundless statements harmful to the institute publicly.

The News article contains other errors. After Manfred Eigen's departure from IMB's supervisory board, for example, the other members did not resign "in sympathy". The only other member who left then had resigned six months previously because of his workload.

Eigen has never been chairman of the supervisory or scientific advisory board of the IMB.

The statement attributed to Eigen — that he was told by the research ministry of the state of Thüringen only to speak to IMB employees in the presence of a ministry representative — is also incorrect. The truth is that the chairman of the supervisory board had expressed his desire to be present at one discussion in June 1996.

Hermann Hamacher

*Thüringen Ministry for Sciences, Research and
Cultural Affairs,
Juri-Gagarin-Ring 158,
99084 Erfurt, Germany*

Gaps in the Human Genome Project

If the multinational Human Genome Project is to continue its successful start, sequencing strategies must be changed.

Jared C. Roach, Andrew F. Siegel, Ger van den Engh, Barbara Trask and Leroy Hood

The adage 'Buy now, pay later' reflects the current operational strategy of the Human Genome Project (HGP). Thirteen per cent of the sequence is now available, and the project seems to be ahead of schedule and under budget¹. But the looming costs associated with gap closure in the final phases of the HGP have yet to be fully addressed. Here we evaluate and quantify those deferred costs.

Currently, the HGP emphasizes highly parallel production of first-draft sequence. 'Highly parallel' refers to the simultaneous initiation or extension of known sequence from many different seed locations. This emphasis represents an acceleration of a subset of the original goals of the project's US arm, stated in 1990, which focused on contiguous high-quality sequence. The current shift in strategy aims for a first draft of 90% of the human genome by the end of next year². Contiguous high-quality sequence has been deferred until 2003. The motivation for this shift initially stemmed from the announcement of a private initiative from Celera Genomics Corporation to compete with the public HGP³. Increased recognition of the public-health importance of raw genomic data, which can facilitate molecular diagnosis and drug discovery, has further stimulated this policy.

The acceleration in sequencing was made possible by improvements in the throughput of automated sequencing machines, both in the number of machines dedicated to the HGP worldwide and in the capacity of each machine. Public capacity is concentrated in five major centres: the UK Sanger Centre in Cambridge, the US Department of Energy's Joint Genome Institute in California and, with funding from the National Institutes of Health, the triumvirate of Washington University in St Louis, the Whitehead Institute at the Massachusetts Institute of Technology and Baylor College of Medicine in Texas. Worldwide capacity, as measured by the amount of high-quality sequence deposited in GenBank, is now 23 megabases per month, and is expected to increase.

The chromosomes of the human genome are 50–250 megabases long, too large to be sequenced directly. They must first be fragmented into intermediate clones about 150 kilobases (kb) long, such as bacterial artificial chromosomes (BACs; see Glossary for a

definition). The resulting collection of fragments, or BAC 'library', compiled from many genomes, is highly redundant. The exact location of each BAC on the target genome is not initially known, but can be found with additional effort and expense. Algorithms for choosing BACs are called BAC selection strategies (Fig. 1).

Improving BAC selection

We show here that the current HGP strategy for BAC selection is inefficient, and we suggest how it can be improved. We evaluate four BAC selection strategies: mapping, random BAC, limited seeding and parking. Following selection, regardless of strategy, each BAC is sequenced by 'shotgun' strategies initially to first-draft quality, and then eventually to high quality, at a total cost of

US\$30,000–40,000 per BAC, although some optimistic estimates of this cost are as low as \$4,000. All these strategies can be implemented if the shotgun sequencing of the BACs is only to first-draft accuracy, but we assume here for simplicity that they are completely sequenced to the high-quality level. Celera's whole-genome pairwise shotgun approach, which does not use a BAC selection strategy, has been discussed elsewhere^{4,5}.

The choice of BAC selection strategy directly affects the overall cost of the HGP. Redundant sequencing of the genome is the main source of inefficiency that can vary with the choice of strategy. Thus, the amount of overlap among selected BACs is a key parameter for evaluating any strategy (Fig. 2). However, the advantages of choosing BACs with minimal overlap must be balanced

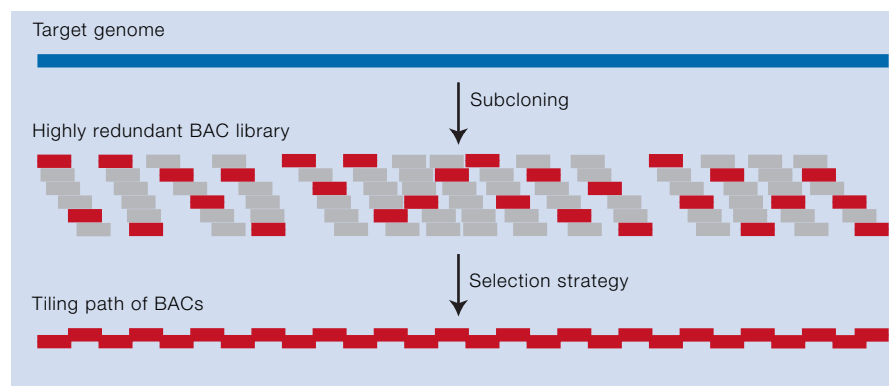


Figure 1 Overview of the strategy for the Human Genome Project. Many replicates of the human genome are fractionated into a highly redundant library of intermediate-size clones (usually BACs). One of several strategies is then used to select a subset of this library for sequencing. In practice, a composite library is used, representative of the genetic diversity of all humans. The actual redundancy of the library is around 30-fold, about ten times that depicted here. (Figure not to scale.)

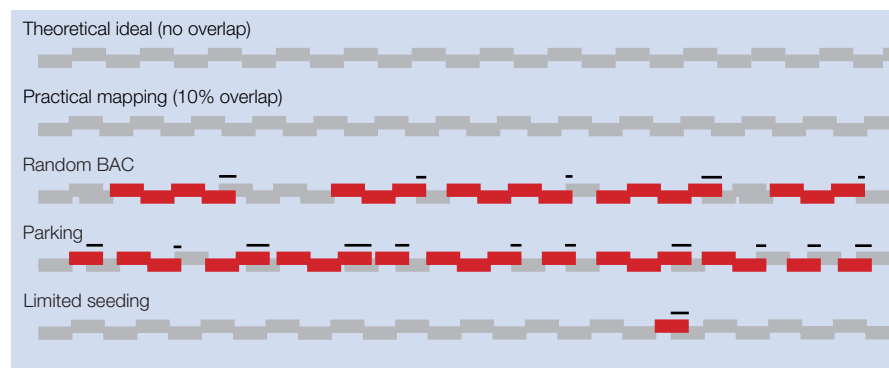


Figure 2 Gap closing for basic genome-sequencing strategies. Grey, BACs sequenced before the initiation of gap closing; red, BACs sequenced for closing the resulting gaps, together with the last BAC filling a gap as part of a limited seeding strategy. Black bars indicate inefficient overlaps caused by mismatches between clone lengths and gap lengths. Implied gap sizes and distributions are roughly to scale.

against the costs of selecting them from the BAC library. Further, the effect of strategy choice on the expense of gap closing must also be considered.

A theoretically ideal strategy would be to completely sequence a minimal tiling path of BACs placed end-to-end across the genome; however, this would be prohibitively expensive (more than \$400 million). Rather, the original strategy for the HGP was to identify a tiling path of BACs with mutual overlaps of about 10%. This up-front mapping approach was more costly and time-consuming than expected and has been abandoned, although mapping data remain valuable for other purposes. The cost of mapping the entire genome with 10% overlapping BACs would be approximately \$240 million⁶.

Without mapping data, BACs must initially be sequenced at random. If no map is then generated from sequence data as they are produced, the strategy is very inefficient. Such a 'random BAC' strategy is not contemplated for the human genome, but we include it here for reference⁷.

A more efficient strategy is to choose a subset of seed BACs, shotgun sequence them and use these sequenced seed BACs to initiate an iterative walking strategy to produce a map. The most likely walking strategy is the sequence-tagged-connector (STC) strategy, which involves selecting BACs that will extend contigs from among a large set of BACs whose ends have been previously sequenced (an STC library)^{8,9}. In practice, the BAC seeds need not be random, as there are enough mapping data to allow about 1,000 evenly spaced seeds to be chosen along the human chromosomes. This combination of limited seed placement with subsequent walking is dubbed a 'limited seeding' strategy.

A variant seeding strategy is to place a very large number of seeds before beginning walking. One such approach screens each seed by limited sequencing (for example, to 0.5-fold redundancy) and rejects the seed if it significantly overlaps any previous seed. This strategy is called a 'parking' strategy, by analogy to cars parked serially along a curb¹⁰.

Gap sizes

Each of these BAC selection strategies produces a different number of expected gaps and a different distribution of gap sizes (Table 1). The parking strategy is the current *de facto* strategy for the HGP, although most centres supplement it with limited regional walking based on previous mapping data. As a result, the HGP now has more than 1,000 seed contigs, with a distribution of gaps between them consistent with either a random or a parking strategy, biased slightly because of the presence of a few deliberately selected seeds. Many HGP statistics are archived at the National Center for Biotechnology Information (<http://ray.nlm.nih.gov/genome/seq>)².

Table 1 Characteristics of basic genome-sequencing strategies

Strategy	Up-front cost (\$ million)*	BACs sequenced to reach 50% coverage	Screening cost (\$ million)	Gaps at 50% coverage	BACs required for gap closing	Overall cost estimate (\$ million)†
Theoretical ideal	>400	10,000	0	NA	10,000	>1,100
Practical mapping	240	11,100	0	NA	11,100	1,017
Random BAC	19	13,900	0	6,900	14,400	1,044
Parking	19	10,000	33	10,000	16,200	1,019
Limited seeding	19	10,500	0	1,000	10,500	759

Numbers are expectations derived from mathematical models and confirmed by simulation^{6,7,10,11}. NA, not applicable.

* See text.

† Estimates for the cost of sequencing a BAC vary from \$4,000 to \$40,000. We use \$35,000 here, plus \$5,000 per gap to account for loss of production-line efficiencies. Library management costs, most significant for the parking strategy, are not included.

The main advantage of the parking strategy is that minimal coordination is required within and between sequencing centres. The only necessary coordination is a central database that can be checked for duplicate effort. Additionally, if the sequencing capacity devoted to the HGP ever exceeded the number of available sites for walking, the additional capacity could be devoted to new seeds rather than sitting idle. One early advantage of the parking strategy was its independence

from any STC library, which meant that BAC selection could be started before the highly redundant STC library became available in mid-1999. A disadvantage of the parking strategy is that few contigs longer than 150 kb are initially produced, which delays analyses requiring longer contigs.

Neither the random BAC nor the parking strategy can be used indefinitely. Once the HGP is half completed, the average random BAC will overlap previous BACs to a

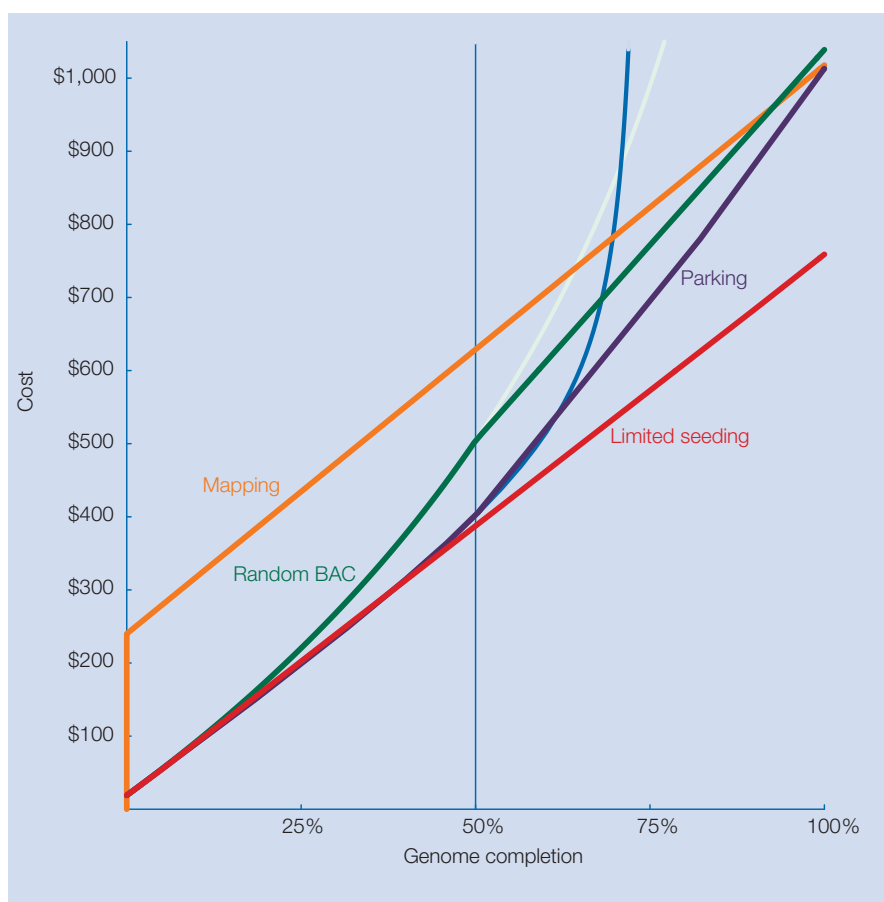


Figure 3 Efficiencies of the basic strategies for sequencing the human genome. Yellow, practical mapping strategy; green, random BAC strategy; red, limited-seeding strategy; blue, parking strategy. Faded curves for the random BAC and parking strategies indicate the consequences of not switching to a gap-closing strategy; solid curves represent a switch at 50% completion. The screening cost for parking is chosen to be the cost of 0.5-fold BAC sequencing. This choice has little effect on our conclusions, as the slope of the parking curve is largely insensitive to the screening cost. The faded curves for random BAC and parking strategies are asymptotic to 100% and Renyi's number (~74.8%), respectively⁹.

prohibitive extent (Fig. 3). For parking, finding non-overlapping seeds becomes increasingly difficult, and eventually impossible at a 'jamming limit' near 75% completion¹⁰. Therefore, after approximately half of these strategies, there must be a switch to a walking strategy. The STC library is therefore needed for all strategies, except mapping, requiring an investment of about \$19 million¹¹.

The cost of closing gaps varies between strategies because there are no perfect mechanisms for choosing clones that exactly match the lengths of gaps. Extremely short gaps can be closed by PCR (polymerase chain reaction), but these are rare. For example, only 5.4% of the gaps are shorter than 10 kb for the parking strategy described here. When gaps exist that are not integer multiples of the effective length of a BAC, a considerable portion of the last (or often the only) BAC used for closing the gap may overlap already known sequence. This source of inefficiency occurs once per gap and thus becomes significant in strategies that result in many gaps. The accumulated inefficiency is the main drawback of the parking strategy, which splinters the unsequenced portion of the genome and maximizes the number of gaps. The amount of inefficiency varies depending on the size of each gap, but averages at about half the length of a BAC per gap.

Costs for each strategy

The mapping and random BAC strategies are more expensive than limited seeding at all stages (Fig. 3). At less than 32% completion, the parking strategy is marginally less expensive than limited seeding, as the screening costs to avoid overlap in parking are initially less than the cost of sequencing overlaps for the seeding strategy. However, even if the parking strategy were to switch to a walking strategy before 32% completion, the final cost of the completed genome would still be greater, because of the larger number of gaps that must be closed. Above 32% completion, the limited seeding strategy is the least expensive strategy at all phases. We estimate the overall cost difference between the parking strategy and the limited seeding strategy to be \$260 million.

Several modifications to the parking strategy can reduce costs. Any reduction in the total number of seeds will reduce the final number of gaps to be closed, improving the parking strategy by making it a form of the limited seeding strategy. Hence, walking based on previous map data or with BACs from an STC library should be implemented for each seed as soon as the seed is produced. Additional seeds should be sequenced only when sequencing capacity exceeds the number of BACs available for walking. Sequencing capacity is currently about 150 BACs per month, and it is unlikely that capacity will exceed 2,000 simultaneous walking steps

Glossary

Bacterial artificial

chromosome (BAC): A device for propagating a small segment of the human genome within an *Escherichia coli* cell, facilitating further analysis. The acronym BAC is used interchangeably to refer to the device itself or to the segment of cloned DNA that it propagates.

Contig: a contiguous tract of known sequence.

Gap: a tract of unknown sequence between adjacent contigs.

First-draft sequence: nearly contiguous sequence with an error rate of about 1:1,000.

High-quality sequence: contiguous sequence with an error rate of less than 1:10,000.

Shotgun: an algorithm that blindly selects clones from a random library for further analysis, such as sequencing. 'Shotgunning' describes the initial creation of a random library.

Mapping: identifying the location of every BAC in a library before sequencing, using methods such as restriction digests or PCR assays.

Random BAC: a tiered shotgun strategy. First, BACs are blindly selected from a genomic library; they are then shotgunned into

smaller plasmid libraries, from which plasmids are blindly selected for sequencing.

Seeding: selection of a set of BACs that will serve as the initiation sites for a walking algorithm.

Parking: a version of the seeding algorithm in which seeds are chosen such that no seed overlaps any previous seed.

Walking: an iterative algorithm for extending a contig. The algorithm 'walks' along the target by using known sequence from the contig as physical or virtual bait to fish for overlapping BACs, then characterizing these BACs so that they may in turn be used as bait.

Sequence-tagged connector (STC): the sequenced end of a BAC (usually produced in pairs). The 'STC strategy' is a walking strategy in which the entire sequence of a contig is used to identify the STCs of overlapping BACs. The minimally overlapping BAC from each end of the contig is then shotgun sequenced, extending the contig.

Minimal tiling path: given a particular strategy, the set of clones with the shortest overall length that spans the genome (see Fig. 1).

Finishing: the final stages of sequencing a BAC. Typically, BACs are shotgunned into plasmid libraries, from which a three- to sevenfold excess of raw sequence data is produced, resulting in first-draft sequence. Finishing eliminates errors and closes gaps in first-draft sequence.

Gap closing: an algorithm for closing large gaps between BACs. By contrast, finishing includes the process of filling in small gaps within the sequence of a particular BAC.

Whole-genome pairwise shotgun: an algorithm for genome sequencing that bypasses a BAC selection strategy. Paired ends of clones of varying sizes are sequenced, producing almost all the target sequence, but fragmented into a very large number of ordered contigs. Finishing would require closing all of the short gaps between these contigs.

Completion: the goal of the Human Genome Project, which includes the complete high-quality sequence of all of the unique DNA in the genome, as well as representative sequences from repetitive DNA. The non-redundant length of the genome is approximately three gigabases.

before the HGP is completed. Innovative methods for tailoring the length of sequence necessary for closing a gap to the length of that gap could also eliminate much inefficiency. PCR is one such method for very short gaps. Constructing an additional STC resource consisting of small insert BACs (for example, 30–100 kb) would reduce certain overlaps, allowing a modest gain in efficiency. Another strategy modification allows small overlaps during the parking screening process. Some modifications are already being implemented. Nevertheless, even with further modifications, the cost differential between limited seeding and parking will remain very high.

The scale of the HGP, as well as its multinational and highly collaborative nature, dictate that a number of political, economic and technical factors influence strategy choice. Our data indicate that some choices may result in false economies. To reduce redundancies, we recommend greater coordination at all levels and that centres agree

on a minimal set of evenly spaced seed BACs. In particular, the current default parking strategy should be abandoned, and the HGP should be directed towards a strategy that reduces the ultimate number of gaps in the genome. ■

The authors are in the Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195, USA. A.F.S. is also in the Departments of Management Science, Finance and Statistics.

Correspondence should be addressed to L.H. e-mail: leehood@u.washington.edu

- Collins, F. S. *et al.* *Science* **282**, 682–689 (1998).
- Jang, W. *et al.* *Trends Genet.* **15**, 284–286 (1999).
- Venter, J. C. *Science* **280**, 1540–1542 (1998).
- Roach, J. C. *et al.* *Genomics* **26**, 345–353 (1995).
- Weber, J. L. & Myers, E. W. *Genome Res.* **7**, 401–409 (1997).
- Siegel, A. F., Roach, J. C., Magness, C., Thayer, E. & van den Engh, G. *J. Comput. Biol.* **5**, 113–126 (1998).
- Roach, J. C. *Genome Res.* **5**, 464–473 (1995).
- Venter, J. C., Smith, H. O. & Hood, L. *Nature* **381**, 364–366 (1996).
- Mahairas, G. G. *et al.* *Proc. Natl Acad. Sci. USA* **96**, 9739–9744 (1999).
- Krapivsky, P. L. *J. Statist. Phys.* **69**, 135–150 (1992).
- Siegel, A. F. *et al.* *Genome Res.* **9**, 297–307 (1999).

Feeling the neurobiological self

The self may be defined in terms of responses to experience.

The Feeling of What Happens: Body and Emotion in the Making of Consciousness

by Antonio Damasio

Harcourt Brace: 1999. 386 pp. \$20, £20

Raymond J. Dolan

The American philosopher Susan Langer, one of the most insightful commentators on emotion, proposed in her book *Mind: An Essay on Human Feeling* (Johns Hopkins University Press, 1988) that the most vexing question in the philosophy of the biological sciences is how something called 'feeling' enters into the physical events that comprise an organism. She suggested that feeling stands "in the midst of that vast biological field which lies between the lowliest organic activities and the rise of mind". Antonio Damasio's stunning book provides us not only with an account of the embodiment of feeling states, but also with a related proposal for understanding two important questions in neuroscience, the nature of self and the nature of consciousness.

From the author's perspective, the problem of consciousness can be described in terms of how the movie in the brain is generated and, more fundamentally, how the brain generates a sense of self that is the very subject of experience. The movie in the brain provides the focus for most current neurobiological work on consciousness; the issue of an experiencing organism is, for the most part, dismissed as currently intractable. Indeed, there is an assumption among neuroscientists that a programme of research focusing on the self runs the risk of ridicule. After all, nobody wants to be accused of resurrecting the homunculus. Yet many of the illnesses that cause the most sustained human suffering, such as schizophrenia, are primarily disorders of selfhood. The astonishing boldness of *The Feeling of What Happens* is that it unashamedly grapples with these issues, and in the process provides the first truly compelling neurobiological account of the self.

Damasio's principal arguments are founded on a premise that the maintenance of a stable internal milieu is an imperative in regulating the life of any organism. Damasio suggests that the emergence, in evolution, of organized neural systems in the service of this function provides the initial steps towards an emergence of self. In effect, the self has a pre-conscious biological precedent, the proto-self, which corresponds to a coherent collection of neural patterns that map, moment by moment, the physical state of the organism.

The relevant neural structures include



Self-portrait: allowed in Chagall's world, but scientists researching the self risk ridicule.

brainstem nuclei, hypothalamus, medial forebrain and insular and somatosensory cortices. This extended system creates a first-order representation of current body states. In parallel, sensory maps are created that represent objects — whether objects in the world or mental images. Their interaction necessarily alters the current state of the organism. This leads to the central proposition that the basis for a conscious, knowing self is a feeling state that arises when organisms represent a non-conscious proto-self in the process of being modified by objects. In essence, a sense of self depends on the creation of a second-order mapping, in certain brain regions, of how the proto-self has been altered.

Damasio proposes important conceptual distinctions between core and autobiographical self, between core and extended consciousness and between emotion and feeling. Core self is the transient protagonist of consciousness that is continuously generated through encounters with objects. In contrast, autobiographical self is heavily dependent on the formation of enduring experiential memories, attention and language. Personal identity is an obvious concomitant of an autobiographical self.

Core consciousness is a consequence of the same mechanisms that generate core self, with the added element of enhanced processing of objects that were involved in generating core self. By contrast, extended consciousness depends on holding in mind, over time, a multiplicity of neural patterns that describe the autobiographical self. An important element here is that these patterns

themselves constitute objects capable of generating core consciousness. What necessarily follows is a functional interdependency between, on the one hand, core self and core consciousness and, on the other, autobiographical self and extended consciousness.

This neurobiological account of core self has an interesting parallel in recent philosophical debates about the nature of self. The philosopher Galen Strawson has proposed that the self could be defined as a subject of experience that is a single mental thing. His introspectively derived account corresponds remarkably closely to Damasio's core-self concept in that both postulate an evanescent self, devoid of identity, which comes into being in response to unique experiences. In Damasio's account, core self presents to the organism a fleeting sense of subjectivity to which knowledge of the moment can be attributed.

Core self and core consciousness have their origins in a mapping of body states and are about two facts, the organism relating to objects and the fact that this relation causes a change in the organism. It follows that a central agenda for a neurobiology of consciousness involves a description of how, in turn, the brain maps the organism and objects, and, most importantly, the dynamics of their relationship.

One of the most important implications of Damasio's thesis is surely that consciousness is not the privileged domain of any sensory system. This is already borne out by observations that acquired damage within a sensory channel does not impinge upon core

consciousness (a possible exception may be the special case of the somatosensory system). Nevertheless, this proposal is likely to be highly controversial in that it casts doubt on any research that seeks to provide a comprehensive account of consciousness from the perspective of a single sensory modality such as vision. Paradoxically, for those of us who study emotion, it is gratifying to think that, without the usual fanfare, we have implicitly been studying consciousness.

At the end of his life, the British poet W. H. Auden movingly acknowledged an indivisibility of self and bodily experience in referring to his body as "my tutor also, but for whose neural instructions I could never acknowledge what is or imagine what is not". To think, imagine and feel are the very stuff of mind and in Damasio's account they are deeply rooted in a sense of body. The exposition of this relatedness in *The Feeling of What Happens* constitutes a remarkable work of intellectual daring. The challenge posed is a radical redefinition of what constitutes the central concerns for a comprehensive account of consciousness. Indeed, by placing human emotion and feeling at centre stage, Damasio ensures their rehabilitation into mainstream neuroscience, a move first initiated in his previous book, *Descartes' Error* (Putnam, 1994). Any of the above achievements would make this book recommended reading; combined, it becomes compulsive and compulsory. ■

Raymond J. Dolan is in the Wellcome Department of Cognitive Neurology, University College London, Queen Square, London WC1N 3BG, UK.

The claim of the inner world

Cassandra's Daughter: A History of Psychoanalysis

by Joseph Schwartz

Penguin: 1999, 339 pp. £20, \$20.95

Jane Kitto

It is a brave man who attempts a history of psychoanalysis in 300 pages. The controversies are legendary, and most people already have an opinion and know something about the personalities involved. We need a history of psychoanalysis for the general reader, but I am afraid this is not it, even though it goes some of the way.

Joseph Schwartz brings Sigmund Freud to life: his early scientific struggles; his search for somewhere to make his mark; his early theories; and the collaboration with Josef Breuer which came unstuck. He deals with Freud's journey towards a workable method, the 'analytic hour'. There is a long section on the vexed matter of the seduction theory and child sexual abuse. Schwartz leaves Freud the theoretician at about 1910 (and about one-

third of the way through the book); the next two chapters describe the early splits with Alfred Adler and Carl Jung.

He then shifts to the United States with a fairly brief account of psychoanalysis there and a long and interesting discussion of the reforms in psychiatric institutions during the early decades of the century, specifically the work of William Alanson White and Harry Stack Sullivan. Most of the last third of the book deals with developments in Britain, beginning with Melanie Klein and child analysis, followed by a brief discussion of Jacques Lacan in France.

This is no potted history. Schwartz began as a physicist, and has become a practising psychotherapist via research into mental health. He is a staunch defender of psychoanalysis, and he handles the question of its scientific status, or otherwise, robustly. It is, he says, "a science in the sense that it attempts to understand human subjectivity in material terms". In the final chapter he takes up the question of neuroscience and psychopharmacology, arguing that although drugs have their place in treatment, they do not replace psychoanalysis; "in the absence of a

deeper understanding of mind-body connections, such pharmacological agents will always have a certain hit or miss quality to them". Neither pharmacology nor psychoanalysis has a magic bullet. Schwartz believes that psychoanalysis is a serious business, and has no truck with dinner-table chatter. The psychoanalytic profession, not having been very good at defending itself, needs this kind of support.

As a scientist, Schwartz has his feet firmly on the ground. But there is another Schwartz who gives the impression of trying to ride two or more horses at once. The book is partial in both senses of the word. While there is no harm in this, provided it is made explicit, Schwartz's preferences are left to the reader to discern — although this is not too difficult. It is not until page 138 that he shows his hand: "This book could be said to be trying to position psychoanalysis as an extension of the Western scientific tradition as a way of understanding the human inner world". But it is sub-titled *A History of Psychoanalysis*, and the result is an uncertainty of focus which leads one to wonder why he has chosen particular topics to expand on.



Capitalizing on psychoanalysis: a psychoanalysing machine exhibited at an industrial show in 1931.

Why so much on psychiatric reform in America? How does it relate to the book's purpose? The choice may appear arbitrary.

Two rather more serious problems are fairness and consistency. Like everyone else, Schwartz has his preferred theorists and they are those who emphasize the external over the internal; for instance, John Bowlby, Donald Winnicott and Heinz Kohut. This seems inconsistent, in view of his description of psychoanalysis as the attempt to "understand the structure and dynamics of the inner world of the experiencing human being". In fact, the question of external and internal is not as simple as Schwartz suggests.

As for fairness, psychoanalysts have from the outset dealt with theoretical disagreements by accusing opponents of personal failings, sometimes verging on character assassination. Schwartz the scientist knows that the theory-maker's personality is irrelevant to the value of the theory, but Schwartz the polemicist cannot resist joining in the fun. Melanie Klein is in danger of emerging as a difficult woman who had some important ideas about children but paid little attention to the outside world. This fails to do justice by a long way to the depth of her ideas, for instance her formulation of the paranoid and depressive positions, which provides psychological grounds for some fundamental questions of society and politics, and their subsequent development by others. By Wilfrid Bion, for example, regarded by many as the most profound psychoanalytic thinker since Freud, but who gets just one mention — in a chapter note.

One must not be unjust to the opposition. To give one example, Joan Riviere was an early and enthusiastic supporter of Klein. She was also a rather conflicted woman and she went into analysis with Ernest Jones, himself an equivocal character, who seems to have behaved quite unprofessionally towards her. Jones finally sent her to Freud, who repaired some of the damage and wrote in forthright terms to Jones about his conduct. Schwartz mentions that Riviere translated some of Freud's papers, but fails to mention that she wrote a number of her own, including one on the clinical problems of narcissism which has become a classic. As a result she emerges as a rather neurotic lady but not as the significant figure she was.

To return to Schwartz's declared purpose, psychoanalysis does indeed have a place in twentieth-century Western tradition, and in his view it is a high destiny. This approach can easily lead to some rather large statements, such as "psychoanalysis has been charged not with participating in the glory of the Industrial Revolution but with clearing up the mess it left behind". One can nod in agreement, or one can take it with a pinch of salt, but this sort of thing is risky, as another example, which may raise some eyebrows this side of the Atlantic, shows — "He went to

Cambridge in 1925, then at the height of its fame in natural sciences".

This could have been a truly well-argued book if Schwartz the scientist had kept a better hold on the other Schwartz. As it is, anyone reading it will certainly learn something about psychoanalysis, and may be stimulated to read further. There is still a book to be written about the development of psychoanalytic thinking and practice which steers clear of the personal. It will be longer and will not be such a racy read. ■

Jane Kitto is an analytical psychotherapist and is at 2 The Park, London N6 4EU, UK.

Dogs, telepathy and quantum mechanics

Dogs That Know When Their Owners Are Coming Home — And Other Unexplained Powers of Animals

by Rupert Sheldrake

Crown: 1999. 323 pp. £16.99, \$25

John Maddox

Rupert Sheldrake is steadfastly incorrigible in the particular sense that he persists in error. That is the chief import of his eighth and latest book. Its main message is that animals, especially dogs, use telepathy in routine communication.

The interest of this case is that the author was a regular scientist, with a Cambridge PhD in biochemistry, until he chose pursuits that stand in relation to science as does alternative medicine to medicine proper. Some readers with long memories may recall that, when Sheldrake's first book appeared in 1981, it was referred to in an injudicious leading article in this journal (*Nature* 293, 245; 1981) under the title "A book for burning?" (where the question-mark was intended as part of the title). The text that followed went on to declare roundly that "even bad books should not be burned", and concluded that the book "should not be burned... but put firmly in its place among the literature of intellectual aberration".

The publicists for Sheldrake's publishers were nevertheless delighted with the piece, using it to suggest that the Establishment (*Nature*) was again up to its old trick of suppressing uncomfortable truths. Many years later, Sheldrake confided to my wife that his children routinely prayed for the soul of the editor of *Nature*, believing that such a wicked person could only come to a bad end.

The motif of the first book, which formed its title, *A New Science of Life: The Hypothesis of Formative Causation*, runs through all the later volumes in this distinctive *oeuvre*. Sheldrake believes that the form or shape of all things, animate or otherwise, is acquired

through the influence of "morphic fields" in a process called "formative causation". Moreover, morphic fields evolve in the course of time, ensuring that when a crystal of copper sulphate or a daffodil first takes on a particular habit, the morphic field ensures that all later crystals of copper sulphate (or all daffodils) follow the same pattern.

In *Dogs That Know ...*, morphic fields have an appendix to themselves, but also frequently recur in the body of the text. In a new twist, Sheldrake cautiously advances the idea that his morphic fields may share with those of quantum mechanics some of the properties of non-locality which offer the chance of making even faster computers some time next century.

Even Sheldrake's fiercest critics will applaud his consistency. The purported role of morphic fields on the shapes of objects in the real world has hardly changed since 1981. It is no less — and no more — than it was then. The idea is borrowed from classical embryology, where a gradient of the concentration of some chemical (such as the protein product of a Hox gene) is supposed to regulate the development of part of an organism's body, and is sometimes referred to as a "morphogenetic field". But morphic fields are evidently all-pervading. No corner of space can be free from them, for then copper sulphate crystals (or daffodils) would acquire different shapes in different locations.

Similarly, Sheldrake's opinion of science has not changed. He speaks of the gulf between "personal experience and the theory that living organisms ... are merely soulless automata", and declares that his experience has made him a holist.

So what is new in this latest book? As Sheldrake describes it, causative formation is no longer merely a hypothesis, but a research programme, complete with the now-standard perquisites of a database and a website (www.Sheldrake.org). And the data? A vast and growing collection of information about incidents involving domesticated animals in interaction with people, or sometimes other animals. The title of the book refers to the copious collections of records, written by people responsible for dogs and, less frequently, cats, describing the pets' capacity to anticipate the arrival home (from work or from a journey) of a second human member of the household. Sometimes the dogs react, not when the second party is leaving the office, say, but when he or she has decided that the time has come to leave.

As an observational programme, none of this is simple. How does a dog signal its anticipation of the homecoming? By waiting at the garden gate, or inside the front door of the house, wagging its tail all the while or showing other signs of excitement. There is the case of a male dog called Jaytee, from a small town in Greater Manchester, who was closely observed and then videotaped in the course

of exercising anticipation of its carer's remote movements. Anticipation was signalled by the dog sitting at a window from which the outside world could be observed. The dog sat in the window on 55 per cent of the occasions when the return journey was under way, but only 5 per cent of the time when the carer (called Pam) was absent. For the videotaped observations, the timing of Pam's return was determined by a third party (using a cellular telephone) at random times: this, Sheldrake says, is when the dog's behaviour most clearly showed the knack of knowing when Pam had decided to return but had not yet embarked on the journey home.

By conceding that the data gathered during these observations are statistically significant, one does not sign up for Sheldrake's interpretation that the underlying mechanism is dog-Homo telepathy. Too many variables are uncontrolled. Did the accuracy of anticipation vary with the length of time elapsed since Pam's departure (suggesting that the dog used its sense of the passage of time to signal its sense of when return was due)? Were there people in the room with the dog (allowing them to communicate somehow with the eager waiter)? And while Jaytee appears to have been chosen for videotaping as a result of his acumen in earlier trials, does not the interpretation of his behaviour require an understanding of the variability of dogs' capacity for anticipation in general? The appendix in which these details are meticulously described is not so much a log of research under way as a record of one of those sets of observations preliminary to the design of properly controlled observations.

The remainder of the main text of this book, meanwhile, is curiously boring. It consists mostly of accounts, running to a few pages at the most, of how horses have been able to find their way home with an injured rider on their back, how cats have cried at the remote death of a human with whom they have been familiar, how dogs have howled when members of the family with which they live have been killed in action in some distant war, and even how people have been taken sick in sympathy with distant injured pets. Especially because people's fondness for their pets often takes the form of projecting onto them human or even superhuman perceptiveness, even more than 1,000 records on the Sheldrake website do not prove telepathy.

I doubt that Sheldrake will take the point. He makes plain his distaste for what he calls orthodox science, which is "all too often equated with a narrow-minded dogmatism that seeks to deny or debunk whatever does not fit in with the mechanistic view of the world". He is habitually courteous and cheerful, but holists of his ilk would not dream of letting controls get in the way of revealed truth. ■

John Maddox is at 9 Pitt Street, London W8 4NX, UK.

Eccentricities of an everyday substance

H₂O: A Biography of Water

by Philip Ball

Weidenfeld & Nicolson: 1999. 387 pp.

£17.99

Frank H. Stillinger

Linking human perceptions of water in mythology, cosmology, politics, literature and the physical and biological sciences (and pseudosciences) may seem to be an idiosyncratic objective. Yet that is just what Philip Ball, until recently a senior editor at *Nature*, has undertaken. With the possible exception of gold, there can be no other substance for which one could write such an engaging account of the profound historical influence exerted on this wide range of subjects. The author's panoramic knowledge, conveyed through a clear and often delightful writing style, makes attractive reading for a technically literate, but not necessarily expert, audience.

The narrative begins with the Big Bang and the subsequent expansion of the Universe that, in due course, produced the elements and chemical compounds, and supplied the early Earth with a chemical and physical ambience conducive to the spontaneous appearance of life. Ball traces our terrestrial geological chronology, and summarizes our present understanding of the role of continental, atmospheric and oceanic water transport in present-day meteorology. And in keeping with the burgeoning interest in planetary exploration in our Solar System, he summarizes the intriguing extraterrestrial evidence for water in both solid and liquid forms that might conceivably have spawned alternative life forms beyond Earth.

Although water has often been seen as 'unique' among liquids, labelling it as such is not especially informative, except in the most trivial sense. Nevertheless, it exhibits an impressive array of anomalies in its physical properties that might qualify it as 'eccentric'. Among these anomalies are the well-known expansion when it freezes at ordinary pressures, and the presence of a liquid-phase density maximum at 4 °C.

The author undertakes to explain, or at least to rationalize, these attributes in terms of the known structure of the water molecule, its resulting electrical asymmetry, and its propensity to engage in tetrahedral arrangements of hydrogen bonds with its own kind as neighbours. The last of these has been amusingly anthropomorphized with line drawings that should appeal to all ages: each water molecule has been rendered as a round-bodied elf whose two arms are destined to grab nearby elfin ankles. This device

is used to introduce the reader to the structure of ordinary hexagonal ice and the forms displayed by snowflakes, and to the structures of high-pressure forms of ice.

In answering the question 'What is liquid water?' by looking at history, Ball has reflected the emergence and maturation of science itself. Beginning with the ancient Greek perception of the substance as one of the primal elements, the shifting answer has reflected specifically the rise of modern structural chemistry. But even in what one might call the 'modern' period (the past half-century, say), considerable revision has occurred to the details of how the liquid is geometrically organized by hydrogen bonding. Imaginative, but now clearly naive, pictures insisted that water should be viewed as modified



Dewdrops: a macroscopic network of molecules linked by frequent but transient hydrogen bonds.

CORBIS/GEORGE LEPP

versions of ice (the 'iceberg' and 'interstitial ice' models), as a 'self clathrate', or as a distorted four-fold coordinated network of hydrogen bonds.

The present consensus seems to be that liquid water is a macroscopic network of molecules connected by frequent but transient hydrogen bonds, which allow unbonded neighbours to occur in numbers that vary with temperature and pressure. Attempts to identify unambiguous patterns of local molecular order that represent portions of the known ice polymorphs have generally proved to be unproductive. In any case, several independent computer simulations for liquid water, using only molecular equations of motion and estimates of the fundamental intermolecular interactions, confirm the random, defective-network viewpoint while automatically producing the characteristic thermodynamic water anomalies.

Quantitative modelling variations on this network viewpoint underlie recent studies directed at an intriguing possibility: does the regime of supercooled liquid water harbour a hidden first-order phase transition between two metastable liquids with different densities (to which the term 'polymorphism' has recently been attached), along with an associated second critical point?

Water that is internal to living organisms differs from its pure bulk form. Aqueous biological fluids are electrolytes, and typically contain a wide array of biopolymers, nutrients and metabolites. As Ball emphasizes, intracellular water is forced to occupy a very crowded neighbourhood indeed, with available channels between the biopolymers being measured in nanometres, if not in ångströms. Consequently, most biological water is surface, or interfacial, water. Yet this aqueous solvent medium has vital roles in controlling the native folding patterns of proteins, and acting as a lubricant for the entire dynamic apparatus of life.

Quantitative details of this solvation role are still a bit hazy, but some broad themes have emerged from research. In particular, Ball provides an account of the present understanding of the 'hydrophobic interaction' phenomenon, which contributes significantly to protein folding and to the structural stability of membranes by driving together hydrocarbon-like molecular moieties.

It is in connection with Ball's discussion of the interfacial properties of liquid water that I would raise a minor quibble with what otherwise appears to be an accurate text. This concerns the comparison of the range of hydrophobic surface effects to the diameter of human hair. By using random samples from my own head as a basis, I estimate that this statement appears to be out by about two orders of magnitude.

As an object of research scrutiny, water has arguably attracted more than its fair share of

questionable, or even absurd, claims. But like it or not, these scraps of pathological science constitute a legitimate part of the biography of water, so Ball provides detailed accounts of three notorious examples: the 'polywater' episode, which began in the former Soviet Union in the 1960s; 'cold fusion', announced by Stanley Pons and Martin Fleischmann in the United States in March 1989; and the alleged homoeopathic phenomenon reported a year earlier by Jacques Benveniste and collaborators for repeatedly diluted solutions of anti-IgE antibodies. Perhaps the inclusion of the last of these was inevitable, given that *Nature* itself was an active but nervous participant in its dissemination.

Some might judge that these and other less prominent water aberrations detract from the scientific legitimacy of the field, but the author intimates correctly that the situation deserves a more positive spin. First, the responses to these challenges to technical common sense affirm the health and vigour of the scientific method. Second, occasional bizarre claims can be interpreted as far-out indicators that creative imagination is widely at play, and this, when suitably filtered, provides the driving force for progress.

I read this book while Hurricane Floyd was inflicting its epic watery damage on the Atlantic coastline of the United States. This was a forceful reminder that understanding, let alone predicting, phenomena of all

scales in our water-rich environment is still woefully inadequate. The mind wanders into musing about what a sequel to this book written at the end of the next century might reveal that this volume cannot. But for now, Ball's contribution is a delightful status report. ■

Frank H. Stillinger is at Bell Laboratories, Lucent Technologies, 600 Mountain Avenue, Murray Hill, New Jersey 07974, USA.

New rhythms of our lives

The 24 Hour Society

by Leon Kreitzman

Profile: 1999. 176 pp. £16.99

Paolo Sassone-Corsi

The moment that Neil Armstrong and Buzz Aldrin took humankind's first steps on the Moon, 30 years ago, characterizes one of the peaks of human technological progress. The event concerned everyone from an intellectual and moral point of view, but the technological impact on our everyday lives appeared minimal. In those days we had no fax, e-mail, Internet or mobile phones. Although we had the capacity to send men walking on the Moon, we were still very slow



Dali on time: we take a more flexible view of the clock than did previous generations.

and inaccurate in establishing long-distance contacts around our own planet. The revolution in communication and computing is a phenomenon of the past few decades. There have been so many dramatic changes in the way we communicate and travel worldwide that modern life has been profoundly affected.

The impetus behind the changes has been 'globalization'; a society has emerged which has an increasing need to communicate throughout the world, at any time of day or night; this is a society that can no longer operate on a 8am–5pm day. By enabling this society to be online 24 hours a day, says Leon Kreitzman, time has been expanded.

Scientists know this only too well. Trying to collaborate with a colleague based in California while in Europe, or arguing back and forth across the Atlantic with a scornful editor over a manuscript were difficult tasks only a few years ago. Time is even more pertinent to the business world, in which ignoring a 24-hour timescale would be financial suicide.

People adjust immediately to new devices, sometimes becoming conditioned to them without even knowing it. We have almost no recollection of how we functioned before faxes and e-mail, and contemporary society would clearly have a hard time functioning without these gadgets.

This is the core of Kreitzman's book, which is entertaining and easy to read, but also refined and comprehensive. Kreitzman ranges from business to informatics, from circadian biology to social science and working regulations. The 24-hour society is most obviously represented by cities in Europe and the United States, but shopping and leisure are also available around the clock in Asia. The demand for flexible timing for these activities has developed in a growing proportion of the population. Of course, this is distinctive of a certain type of social economy, as Karl Marx (cited by Kreitzman) rightly reasoned: "labour during 24 hours of the day is the inherent tendency of the capitalist production."

Working mothers are a particularly apposite case, as most of them need to have a more stringent organization of their time because a high proportion of partners still do not help with housekeeping and child care. They are a group of people who can greatly benefit from a society that allows flexible use of facilities and services.

Another interesting example is the tourism industry, which has grown dramatically in the past couple of decades. Today tourism is global; in the 1980s it took over from oil as the world's largest industry, and now today it accounts for 10 per cent of the world's gross national product. People like to travel far afield, and as their time is limited, they want to have services, tourist attractions, museums, shopping and

dining available at all times, everywhere they go.

Humans, distinct from all other animals, have had their circadian rhythms modified by social habits, electric light, television and international travel. Our natural sleep/wake cycle, as well as our endogenous biological clock, are continuously challenged by external stimuli that have nothing to do with the natural day/night astronomical cycle. Kreitzman gives a remarkably simple but clear (and correct) synthesis of our present knowledge of the biological clock. His hope is that "once we can control our rhythms many of the objections to shift-working and to the 24 Hour Society will fall away". The intrinsic danger of such thoughts is the possibility of biological control over people by society. But Kreitzman is careful to emphasize that advantages of the 24-hour society should not be "achieved by the exploitation of the health and safety of groups in the population".

So, is the 24-hour society some kind of present/future positive condition, or just a nightmare in which the Western world is caught up? Our parents say that it was so much better in their day, when the pace of life was slower and they had more time to enjoy valuable moments. Today it is difficult to have a normal conversation with a friend or colleague without being interrupted by a beep or a ring, and it does seem to many that enjoyable moments are becoming rarer. But I feel that Kreitzman is right; it is just a matter of getting adjusted to new rules and redefining temporal relationships. Then we will gradually, but continuously, learn to enjoy a different timing for the important things in our lives, and by the same token possibly gain time. ■

Paolo Sassone-Corsi is at the Institut de Génétique et de Biologie Moléculaire et Cellulaire, BP 163, 67404 Illkirch, Strasbourg, France.

Know thyself genetically

Genome: The Autobiography of a Species in 23 Chapters

by Matt Ridley

Fourth Estate: 1999, 752 pp. £18.99, \$26

W. F. Bodmer

Genetics, I believe, will stand alongside computing and nuclear physics as one of the outstanding areas of development by which this century will be remembered. The century started precisely with the re-discovery of Mendelism and it will finish with a first draft of the complete DNA sequence of the human genome. That remarkable catalogue of the human genes, ordered along the chromosomes to give us

the Book of Man, will form the basis of most future biological and biochemical investigations of humans. The end of the Human Genome Project is the beginning of the real genetics of mankind. And this is the substantial challenge for understanding over the next century or more.

By combining the power of molecular genetics with classical population genetics and quantitative analysis, new approaches have been devised, not only for the 'positional' cloning of Mendelian traits whose biochemical basis was totally unknown, but also for identifying genes that contribute to the inherited susceptibility to multifactorial traits and diseases. Genes associated with nearly every Mendelian disease have been cloned, however rare the disease. But recall the great physician William Harvey, who urged us to treasure our exceptions, from which, as we now know, so much can be learnt. There is an exciting story to be told about genetics in the genomic era, but that is not to be found in Matt Ridley's book.

His title leads to the expectation that these exciting developments in genetics will be revealed to the general reader. However, in the preface we are told that "this is not a book about the human genome project", but instead "a book about what the project has found". The book is based on an intriguing idea, namely, to progress through the chromosomes chapter by chapter, using in each an example to illustrate what the genome reveals when it is properly read and interpreted. Ridley, a professional science writer and journalist, has, however, followed a common journalist's pathway and selected a collection of topics that he presumes will interest the reader whether or not they are relevant to his professed goal. Too many of these topics, such as sexual evolution, IQ, personality, stress and its relationship to immunity, ageing, memory and even, finally, free will, are simply not yet ready to yield to modern genetic approaches. They are mostly a basis for largely unsubstantiated speculation described in a way that could have been written without any reference to what the genome has to tell us.

Chromosome 1, the largest, is, according to Ridley, apparently empty, although a discussion of the rhesus blood groups might have been of interest. Chromosome 2 is used only as a vehicle for comparing *Homo sapiens* with the chimpanzee, the first of many "mind boggling" comparisons, in this case because of Ridley's apparent surprise that it may take only a relatively small number of differences in the DNA code to distinguish man from chimpanzees. Huntington's disease occupies chromosome 4. Certainly, it is an interesting story, but hardly, as the author claims, the most talked about genetic disease.

Given my own long-standing involvement with the HLA system and the genetics



Chapter and verse: the entire contents of the 23 human chromosomes should soon be known.

of the immune response, such topics would have been my natural choice for chromosome 6. But immunology, in spite of its enormous interest and dramatic development, hardly gets a mention, even in the discussion of asthma, and the only reference to the HLA system is in relation to the rather questionable data on mating preferences. Ridley's chromosome 6 is devoted to the somewhat flimsy evidence for a gene with a possible influence on IQ.

There are some good chapters. The relationship between the apoE lipoprotein gene and Alzheimer's disease is well told and the chapter on BSE is also good. Ridley even mentions James Parry, the geneticist in Oxford whose sheep pedigrees told the correct story about scrapie, as he explained to me when I first came to Oxford in 1970, and which no one believed. A good chapter on eugenics is seriously marred by the fact that it draws the analogy between screening for Down's syndrome and Nazism. Even negative eugenics is hardly applicable to Down's syndrome, and the importance of doing genetic screening only if benefit can be derived from it is hardly mentioned.

Historical treatments of scientific discoveries are a good vehicle for presenting the underlying science, but not when they are as abbreviated and distorted as they often are in *Genome*. There were, for example, several co-discoverers of the *p53* gene; and the molecular biology of prions was uncovered by Charles Weissmann and his colleagues. The story of Gregor Mendel's discoveries is described as though they appeared from nowhere, and ignores the evidence that they were stimulated by his abbot's interest in uncovering the basis for the animal and plant breeding used in the monastery's farms. It was the theoretical physicist

George Gamow, not Francis Crick, who first suggested codes relating DNA to protein. I still remember when, as a young mathematical geneticist, I sought Crick's advice about problems I might investigate in molecular biology, and he suggested that I explore the mathematics of overlapping codes. Fortunately and correctly, I did not see how my mathematics could help to provide a solution to the coding problem, which eventually came entirely from experimental approaches.

Style is, I suppose, a matter of personal preference. "Now I tell you again Dear Reader" somehow does not, to my mind, match the topic of 'genome'. Suddenly, towards the middle of the book, we are told "the human genome project is founded upon a fallacy"! He has clearly not yet explained the importance of human variation, which underlies any explanation, for example, of the balance between nature and nurture. Perhaps it is this he is referring to when he says, "the tension between universal characteristics of the human race and particular features of individuals is what the genome is all about". There is also the occasional strange, throw-away sentence: "But then judges were never very good at science" seems a remarkably injudicious comment, and I wonder on what basis he makes this extraordinary suggestion.

Ridley has written well in the past on evolutionary topics, and is clearly capable of writing in a stimulating way for the general reader. Perhaps in future he would do best to stick to topics involving evolutionary ideas, where he is most at home. If this book kindles an interest in the genome and what can be found out using modern genetic approaches, then perhaps it is doing a service. But if you want a serious and challenging discussion of

what genetics and the genome can tell us about humans, I am afraid you must look elsewhere.

Walter Bodmer is at the Cancer and Immunogenetics Laboratory, Imperial Cancer Research Fund, Institute of Molecular Medicine, Oxford OX3 9DS, UK.

Darwin's evolution

Almost Like a Whale: The Origin of Species, Updated

by Steve Jones

Doubleday: 1999. 402 pp. £20

Mark Pagel

Gilbert White, writing in 1773 in his *Natural History and Antiquities of Selborne*, understood why owls have unusually "soft and pliant" feathers on the tips of their wings, but he could not begin to comprehend why "some females of the brute creation ... devour their young", adding "I leave [this] to abler philosophers than myself to determine". Writing a century later, a biologist gave the answer, "these practices appear to have originated in savages recognising the difficulty, or rather the impossibility of supporting all the infants that are born".

The biologist was Charles Darwin, whose *Origin of Species* had by the time of this quote already earned him fame and opprobrium in various quarters. The reason was not that Darwin could explain better than others the functions of the diverse traits and characteristics of animals, but that he had offered a mechanistic account of how animals, including humans, come by those traits. That mechanism he called 'natural selection', and so radical were Darwin's views that much of the *Origin* can be seen as a highly self-conscious defence of his claims for its near-exclusive role in shaping life. Every page carries a sense of Darwin's disquiet that, at least for the reader, some insuperable challenge to natural selection will arise: Darwin repeatedly reminds us of the truth of "my theory", or the plausibility of "my account".

By all reasonable measures, the theory has given a good account of itself. Academic journals, books, newspapers, radio and television brim with 'darwinian' (this journal now insists on the lower case in the adjective) investigations of this or that, and they are not limited to biology. Economics, politics, philosophy, linguistics, psychology, medicine and other fields have succumbed to the darwinian juggernaut. Novelists are asked to debate its merits. Sophistication shows up where least expected: the Kansas Board of Education creationists who recently banned Darwin from Kansas schools accept the reality of microevolution (evolution within species); they just don't like macroevolution



Almost like a whale: the hippopotamus is the whale's closest living relative.

(one species descending from another). Indeed, so conventional has Darwin become that even the Pope accepts the reality of evolution, although he shrewdly insists that souls are outside its reach.

Steve Jones equally shrewdly realizes that all this makes *Origin of Species* a good candidate for the book of the millennium — a view others share. It may come as a surprise, then, to learn that he has chosen to rewrite it. Is this the scientific rationalist's variation on blasphemy? Jones thinks not, suggesting that, although no one ever put the case for 'evolution' better than Darwin (ironically, Darwin avoided that word, using it only once in the *Origin*), the facts in the *Origin* "are those of a century and a half ago and leave many gaps before its case can be considered proven". Reverently, Jones retains and uses as his own framework Darwin's chapter titles, chapter summaries, lists of chapter contents and, curiously, all of Darwin's final chapter. The title is his own, although it, too, derives from a line in the *Origin*.

Jones's opening remarks cite the familiar refrain of the "100 million Americans" who believe that "God created man pretty much in his present form at one time during the last ten thousand years". In fact, even this represents progress for the theory: these same 100 million people have been quoted since I was a boy, when they made up nearly 50 per cent of the US population. Now there are something like 280 million Americans.

Most of the book's introductory chapter, however, is devoted to the AIDS virus. Jones observes that "the human immunodeficiency virus contains in its brief history the entire argument of *The Origin of Species*". And so it does. Research on HIV reveals its struggle for existence, the action of natural selection, its

geographical spread and descent with modification. How much more about evolution and adaptation must be known if other organisms are included? In succeeding chapters, Jones moves on to these other organisms, and draws on a prodigious inventory of facts to support Darwin's original musings on 'variation under domestication', 'difficulties of the theory', the 'imperfection of the geological record', 'hybridism' and more.

Imagine Darwin's delight had he known that hippopotamuses are the closest living relatives to the whales, or that Arctic and Antarctic fishes have independently evolved anti-freeze, or that some insects got their wings from the gills of other insects. Where Darwin, of course, had no knowledge of or access to genetic information, Jones reports many of the latest charms of evolutionary genetics, such as horizontal gene transfer, selfish DNA and genome mapping projects. All this is reported in his staccato style, suitable for the general reader and yet informative for the specialist. He is seldom more than a paragraph away from some amusing or thought-provoking anecdote: the allegations about George Spencer and a one-eyed pig are memorable.

Jones is sceptical about much of the work that applies principles of natural selection to understanding human social evolution and behaviour: human sociobiology. Darwin, whose *The Descent of Man — and Selection in Relation to Sex* remains startlingly bold and perspicacious in doing just that, would not, I think, approve. True, some human sociobiological studies border on the crude and offensive. Others, however, are producing some of the most challenging and stimulating work in evolutionary biology. Humans are to some extent the final frontier, always until now pro-

tected from the vulgar apparatus of biological evolution by intelligence and cultural innovation. But demonstrations that cultural practices, including wealth inheritance and marital systems, family size and differential investment in children, are explicable on darwinian grounds have demolished the view of humans as above the fray.

Nevertheless, Jones remains one of our most effective commentators and writers on science. *Almost Like a Whale*, at around 400 thick pages, is almost like a whale to carry around. But it repays the effort. This 'Origin of Species — Revised Edition' may lack the sense of danger and discovery of its predecessor, and the radicalism has been replaced by a presumption of the truth of natural selection. Yet Jones succeeds impressively in his desire to bring us up to date on the facts of evolution: to read this book is to see in one place much of the sweep and grandeur of evolution by natural selection. ■

Mark Pagel is at the School of Animal and Microbial Sciences, University of Reading, Whiteknights, Reading RG6 6AJ, UK.

Coelacanth à la Marseillaise

A Fish Caught in Time: The Search for the Coelacanth

by Samantha Weinberg

Fourth Estate: 1999. 239 pp. £13.99

Philippe Janvier

In December 1938, the first living coelacanth (*Latimeria chalumnae*) to become known to science was accidentally caught near East London in South Africa. Coelacanths were thought to have been extinct for 70 million years — since the late Cretaceous — hence the sensational impact of this discovery of a 'living fossil'. But, in addition to this astonishing survival, the coelacanth revealed the anatomy of the soft tissues of 'crossopterygian' fishes, thought to have been forerunners of the four-legged vertebrates, or tetrapods. Therefore, it raised considerable interest among evolutionary biologists and palaeontologists. Samantha Weinberg's book describes the history of the discovery of the living coelacanth, and the lives of the women and men involved in it.

Several books have been published previously on the story of the Comoran coelacanth, including *Old Fourlegs* by J. L. B. Smith (Longmans, 1956), who first described the fish and named it after its discoverer, Marjorie Courtenay-Latimer. But Weinberg's book also includes the account of the recent discovery of the Indonesian coelacanth. In her book, both discoveries sound like detec-

tive stories, with obscure political and ethical intrigues, often involving French scientists.

The first coelacanth was mounted in such a way that little of its anatomy was available to Smith, who was eagerly waiting for a second specimen. Thanks to Eric Hunt, the second coelacanth turned up in 1952, but in the Comoros. The way it was taken from the Comoros, then a French colony, by Smith himself, on board a plane chartered by South Africa's prime minister, Daniel Malan, raised frustration among French scientists and led the French government to ban foreign coelacanth investigators. Subsequently, most coelacanths caught in the Comoros were sent to Paris and then either used for anatomical study or donated to major scientific institutions, with the stipulation that they would not be dissected.

Smith accused the French of hindering the scientific study of this extraordinary fish, and appealed, in vain, to an international research programme. Weinberg clearly sympathizes with Smith on this matter. Alas, generosity is rare when a country possesses an invaluable research subject. In 1952, French science was recovering from the war and the sudden discovery of the coelacanth in French waters could not possibly be ignored. Whatever judgement one may pass on the appropriation of the coelacanth by the French, or, conversely, on the cavalier expedition of Hunt and Smith, the resulting monographs on the anatomy of *Latimeria*, by J. Millot and J. Anthony (and D. Robineau for the last volume), remain a masterpiece. Nevertheless, many French zoologists felt that Smith should have been closely involved in the anatomical study of the fish. After all, neither Millot

nor Anthony were initially ichthyologists, whereas Smith was in 1952, at any rate.

The discoverer of the Indonesian coelacanth, Mark Erdmann, had a radically different attitude from Smith's, as he immediately donated the first specimens to an Indonesian institution and established an agreement for a US-Indonesian joint study of the specimen, in particular a molecular sequence study to check whether it is the same species as the Comoran coelacanth. All went smoothly, and the discovery of the Indonesian coelacanth appeared in *Nature* (395, 335; 1998) in September 1998. Mark and Arnaz Erdmann were heroes, as Courtenay-Latimer and Smith had been in 1939, but with, in addition, a fairy-tale atmosphere because this latest discovery was made during the young couple's honeymoon.

Then, in April 1999, came the shock: the French, "inevitably" (in Weinberg's words), published a paper aimed at giving a scientific name to the Indonesian coelacanth (*L. menadoensis*, after Menadotua Island, where it was caught) on the basis of a quick morphological and molecular sequence analysis that suggested some difference from *L. chalumnae*. In fact, the authors of this note are both French and Indonesian, with Laurent Pouyaud (not Pouyard) as the first and only French author. Apparently, the Indonesian authorities gave them free access to samples from the specimen donated by the Erdmanns. To play devil's advocate, one may say that the Indonesians did exactly what Smith wanted the French to do. Unfortunately, this new page of the eventful history of the coelacanth came too late to be dealt with in detail by Weinberg, and is only accounted for in a justifiably angry footnote.

The naming of the Indonesian coelacanth is at odds with the ethics of systematics: no reference to the discoverer of the specimen (notoriously studying it), no designation of the holotype (admittedly, a single specimen to date), no museum number and, above all, a wide spreading of the new name in the press before the scientific publication appeared. Immediately, heated debates occurred among systematists as to whether the name was acceptable, or should have the authorship of the first journalist who published it in a daily newspaper, instead of that of Pouyaud and his co-authors. But, according to the requirements of the current edition of the International Code of Zoological Nomenclature, there seems to be no reason to reject it and its authorship, whatever its ethically questionable background. Further studies, however, may prove *L. menadoensis* to be a mere synonym of *L. chalumnae*. The close collaboration between the German Hans Fricke and the French Raphael Plante in the first successful observations of the Comoran coelacanth alive and well in its natural environment contrasts greatly with the history of the naming of the Indonesian one.

Weinberg explains why the coelacanth is interesting and at the same time disappointing, in the context of the question of the origin of tetrapods. She also gives an account of the discovery of the lungfishes, in the nineteenth century, which raised the same excitement as that of *Latimeria*. However, this section and the section about fossils contain a few minor mistakes (for example, the spiral valve is not in the stomach, but in the intestine). The coelacanth is currently classified among the sarcopterygians (vertebrates in which the paired fin/limbs skeleton articulates to the girdles by means of a single bone, the humerus and the femur in tetrapods). Along with lungfishes, the tetrapods and several fossil-fish groups, it shares very few advanced characters with the tetrapods, and this puts it somewhere near the base of the sarcopterygian tree. In a sense, the coelacanth tells us more about the primitive condition of all bony fishes than about the origin of tetrapods.

As a whole, the book reads well, like an adventure story, with amusing and charming anecdotes about the lives of the protagonists. To a scientist, it may be frustrating not to read more about how palaeontologists reacted to the discovery of *Latimeria* and how it could test their previous reconstructions. Although, admittedly, most palaeontologists who dealt with coelacanths and other fossil sarcopterygians before 1938 and experienced the pre- and post-*Latimeria* times are now dead, very few have written about this.

The conclusion of the book somewhat tails off. It mentions the possibility of finding more coelacanths elsewhere in the world, which is plausible, and the need to protect and "leave the coelacanth in peace", which is



Resting in peace: Smith's hand lies on the head of the second coelacanth found off the Comoro Islands.

wise. Nevertheless, the author should perhaps have questioned scientists about what more they expect from this fish. Its behaviour is barely known and, above all, its embryonic development is totally unknown. Knowing how the strangely articulated braincase or the fins of the coelacanth develop would be just as important as what we already know of its anatomy, in particular for understanding the early evolution of bony fishes and sarcopterygians. This is perhaps the next challenge for the forthcoming century.

Lay readers will certainly enjoy *A Fish Caught in Time*, but it can also be recommended to students and zoologists who are too young to have experienced this extraordinary zoological adventure. ■

Philippe Janvier is at the Laboratoire de Paléontologie, Muséum National d'Histoire Naturelle, 8 rue Buffon, 75005 Paris, France.

Pogo-centricity

Ants at Work: How an Insect Society is Organized

by Deborah Gordon

Free Press: 1999. 208 pp. \$25

Jürgen Heinze

It is certainly an odd sight in the early light of an Arizona desert morning to see a number of people digging themselves into the dry soil, crawling around with their noses close to the ground, hunting for something that is invisible from even a short distance. Still, the inhabitants of Paradise and Portal in southeasternmost Arizona are by now presumably quite used to it, because over the past decades the Chiricahua Mountains and adjacent San Simon Valley have become a favourite field site for researchers interested in the evolution, sociobiology and ecology of ants. One of them, Stanford professor Deborah Gordon, now writes about what she has discovered, in 17 years of observations, lab and field experiments and computer modelling, about *Pogonomyrmex* harvester ants.

Ants at Work is an entertaining mixture of personal travel report and popular science. And it is deliberately 'Pogo-centric', as the author herself says: it focuses almost exclusively on this genus, ignoring most other ants. It is also much more a private, 'Gordon-centric' *oeuvre* than a comprehensive and integrated treatise on ants, and three-quarters of the roughly 45 references in the appendix are by the author and her research team. Other research on *Pogonomyrmex*, including studies done just across the road from her own study site and a recently published book on harvester-ant biology, is left aside, as are other key publications, including several fundamental reviews on ant biology. These omissions might perhaps help a naive reader, whose knowledge does not go



Mating time: ants display enormous variability in reproductive tactics across the 300 ant genera.

much beyond the film *Antz*, to keep on track and not go astray among the fascinatingly complex universe of social insects.

Those ant novices will be initiated straight to *Pogonomyrmex*, their foraging strategies, their conflicts with neighbouring colonies and how they make decisions, without the detour to other ants with alternative ways of life. And they will also learn a lot about how all this information is gathered and how researchers cope with the manifold problems that generally make field studies on animal behaviour quite tricky — such as temporal changes in activity patterns or motivation. *Ants at Work* vividly involves the reader in several ingenious experiments especially designed to overcome these and many other difficulties.

Readers more familiar with ants, however, will occasionally feel somewhat uneasy. *Pogonomyrmex* is only one of about 300 ant genera and it has a rather specialized ecology and behaviour. Generalizations, such as those made in the introduction that workers of all ants are sterile and new reproductives fly off to mate, unduly oversimplify the enormous variability of reproductive tactics in ants, with wingless sexuals mating in the nest or workers laying both fertilized and unfertilized eggs.

Furthermore, some aspects of ant biology presented to the reader as amazing news have

for a long time and in much detail been known from other ants. It is common knowledge among social insect researchers, for example, that an ant colony operates without any central control, that workers can easily switch from one task to another, and that the meaning of signals used in communication is context-specific. That an ant queen is not in charge, claimed on the book jacket to be a revolution in our thinking on natural organization, is therefore trivial. On the other hand, the remark that no ant has power over another is somewhat puzzling, given the accumulating evidence for overt kin conflict and dominance hierarchies in some ant species.

One of the book's major themes is division of labour, a very acute problem that is being widely (and heatedly) discussed among social-insect researchers and has also elicited considerable interest in the artificial-intelligence community. Here, Gordon describes a series of experiments and observations which demonstrate that task allocation is highly flexible and that encounter rates between individuals may be important for task switching in *Pogonomyrmex*. She then uses this evidence to shatter the alternative hypothesis that an ant's task is innate, that is, that a worker does a single task throughout its life. However, to my knowledge, this latter viewpoint has never been

held seriously, not even by George F. Oster and Edward O. Wilson, to whom it is attributed. This looks to me like fighting a straw man. I would have preferred an up-to-date overview on the state-of-art in this field or a thorough analysis of how Gordon's encounter-rate model compares with temporal polyethism (which is mentioned elsewhere in the book) or the many other models of decentralized control and self-organized division of labour. Even readers less familiar with social insects would not have been overtaxed by a balanced treatment of this subject.

To conclude, *Ants at Work* offers a lot of information on harvester ants and introduces several new, and some not so new, interpretations of ant behaviour, including the concept of self-organization, to a broader audience. This is probably the author's main aim, and it is well achieved. It is doubtful, however, whether the book will satisfy the much farther-reaching aims listed on the dust-jacket. I enjoyed reading the book, but it did not really revolutionize my understanding of ants. ■

Jürgen Heinze is at the Zoologisches Institut I, Universität Erlangen–Nürnberg, Staudtstrasse 5, D-91058 Erlangen, Germany.

An exobiologist's life search

Carl Sagan: A Life in the Cosmos

by William Poundstone
Holt: 1999. 560 pp. \$30

Carl Sagan: A Life

by Keay Davidson
Wiley: 1999. 540 pp. \$30

Christopher Chyba

The astronomer Carl Sagan died of pneumonia in December 1996, following a two-year struggle with the bone-marrow disease myelodysplasia.

Sagan had directed the Laboratory for Planetary Studies at Cornell University, where his research group pursued questions related to the organic chemistry of the outer

Solar System, the origins of life on Earth and the search for life elsewhere.

Always optimistic about the prospects for extraterrestrial life and intelligence, Sagan's early scientific contributions nevertheless considerably diminished hopes for life on Venus and Mars. His subsequent work on the "early faint Sun paradox" set the stage for decades of research trying to reconcile the evolution of the Sun with the early terrestrial greenhouse atmosphere.

Throughout his career, he played a leading role in Solar System exploration, including the Viking and Voyager missions. Sagan's popular book *Cosmos* (Abacus, 1983), based on his award-winning television series, was the best-selling science book ever published in the English language. At the time of his death, he was undoubtedly the most famous scientist in the United States. He was a vocal enthusiast for the scientific search for extraterrestrial life, a public sceptic about claims for the paranormal and UFO visitations, and had played a prominent public role in the 'nuclear winter' debate, warning of the possible effects that the fires resulting from nuclear war would have on climate. This was a remarkable brew, and guaranteed that his life would prove of unusual interest. His first two biographies have just been published.

Carl Sagan: A Life in the Cosmos is by William Poundstone. His previous book, *Prisoner's Dilemma* (Doubleday, 1993), was part biography of the mathematician John von Neumann, part popular explanation of game theory and part cold-war history. As a result, Poundstone brings to Sagan's life a feeling for science combined with an intuition for politics and bureaucracy.

Poundstone understands the psychological free-for-all at work when scientists choose which problems to pursue: "These decisions are necessarily intuitive and idiosyncratic, based on incomplete information and personal value judgments. A prime example of this is Sagan and the exobiologists' belief that, of all the possible things we might learn about the universe, the detection of extraterrestrial life would hold unique significance." But he also understands Karl Popper's distinction between this personal realm of hypothesis formation and the intersubjective, though tricky, realm of hypothesis testing. This gives Poundstone's accounts of both exobiology and the nuclear winter controversy some sophistication. About nuclear winter, he writes, "the TTAPS [Richard Turco, O. Brian Toon, Thomas Ackerman, James Pollack and Sagan] group was qualitatively right and quantitatively wrong in their first publication — and then they revised as new data became available. What makes nuclear winter different is that the error-correcting mechanism of science was played out in the political arena."

The second new biography is *Carl Sagan: A Life*, written by Keay Davidson, a science writer for the *San Francisco Examiner* newspaper. Both biographies contain a great deal that is unflattering to the young Sagan. Each particularly criticizes Sagan's behaviour in his first marriage to Lynn Alexander (now Lynn Margulis, the biologist). Sagan was 22 and Alexander 19 when they married. Clearly, the union was premature. "The divorce was probably a greater success than the marriage," Poundstone comments dryly.

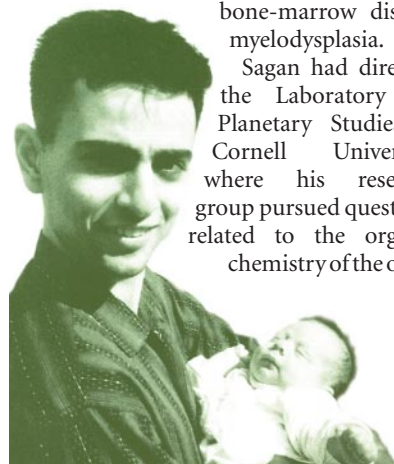
Neither biographer can be blamed for quoting a divorced spouse, a bitter child of a first marriage or even a long-time friend who chooses to reveal old confidences. Poundstone is interested in Sagan's personal life, but lets those he quotes speak for themselves. Davidson's biography, by contrast, is replete with attributions of motive and unsubstantiated accusations that go well beyond the evidence.

Consider Davidson's description of the teenage Sagan at the University of Chicago (Sagan went to university at 16): "He apparently did not show special favors to women — did not talk down to them, as if they were children — which, one surmises, at least some of them found strange but alluring, like being seduced by someone eager to ravage a previously unravaged part of their bodies: their brains."

"Whether he ravaged anything else is unclear; Sagan left virtually no reminiscences about his college dating habits, and what he did recall was typically self-centered, concerned less with the girl than with how he thought she perceived him."

It is hard to know what to make of this disturbing passage. Quite apart from his strange "surmises" about women, Davidson criticizes Sagan for leaving no reminiscences about his romances with the "girls" he met. No honourable explanation is entertained, such as simple respect for the young women's privacy.

More serious is Davidson's accusation that the young Sagan wilfully and illegally revealed classified information in order to further his chances of winning a Miller Institute graduate fellowship to Berkeley. This is a serious and specific legal allegation which Davidson does not substantiate. Instead, we learn that Sagan was nominated for the Miller fellowship by Nobelist Joshua Lederberg, that the Nobelist H. J. Muller wrote in support that Sagan "stood head and shoulders above" other students of his age and that he had recommendations from the future Nobelist Melvin Calvin and the chair of Berkeley's astronomy department. Yet Davidson claims that these alone were "probably not good enough to ensure his victory" in the fellowship competition. Therefore, he asserts, Sagan revealed the classified information to "make the Miller judges sit up and take notice".



From inner to outer space: Sagan in 1959 holding his and Lynn Margulis' son, Dorion.

Davidson writes: “[Sagan] decided to confide to [the fellowship judges] information that he was required by federal law to keep secret. He revealed his research at the Armour Research Foundation on the remote detection of lunar nuclear explosions. He must have known the risk he was taking. The information was classified; he had previously cautioned Muller not to discuss it with others. ... Washington did not look kindly on the leaking of nuclear information.” The only footnote to this passage reads “Sagan’s Miller Fellowship file”.

Davidson presents Sagan’s request to Muller not to discuss his work as evidence that Sagan knew he should not be revealing it. Davidson faxed me the letter on which he said he based this account: a letter from Sagan to Muller, dated 18 March 1959. The relevant section reads:

“Since [my work on possible primitive lunar organic synthesis] is only a first draft, and since part of it was done on a classified contract — although the enclosed of course is not classified — I would not want extensive distribution of the manuscript at the present time.”

Sagan was asking Muller not to distribute an unpublished manuscript, a standard request between scientists. The manuscript was not classified, and had already been sent to five other scientists. The letter, which was Davidson’s source, provides no evidence that Sagan was revealing classified information.

What information did Sagan “reveal” in his application? The only relevant part of Sagan’s fellowship application (which I examined at the Miller Institute) is the section in which he cites his publications, where he mentions: “Various classified papers on scientific information to be gained from possible lunar nuclear detonations, Physics Research Department, Armour Research Foundation, Chicago.”

But classified documents may well have unclassified titles or subsections. To level the grave charge that someone illegally revealed classified nuclear information, a reporter has a responsibility to provide evidence for the allegation. Davidson does not.

In the early 1990s, the National Academy of Sciences denied membership to Sagan.



A man for all seasons: Sagan speaking as an anti-nuclear advocate at the Great Peace March in 1986.

Both Poundstone and Davidson shed new light on the heated debate that occurred over this affair. Based on sources in the academy, some of whom requested anonymity, it becomes clear that Sagan’s opponents felt the greatest contempt for his popularization of science. There was also an accusation that Sagan failed in his Venus greenhouse work to credit Rupert Wildt, who first proposed the idea. But, as Davidson notes, the opening sentence of the introduction to *Venus* in Sagan’s thesis lists Wildt by name, acknowledges his precedence and cites Wildt’s 1940 paper.

Lynn Margulis, a participant in the academy’s debate, later wrote to Sagan, saying that his opponents were “jealous of your communication skills, charm, good looks, and outspoken attitude”. Poundstone tells us that Sagan guessed the topic of Margulis’ letter and asked his wife, Ann Druyan, to read it to him, omitting names. This is consistent with Carl’s account to me; he preferred not to know the names of those who spoke against him because that could make it harder to

work with them in the future. When giving me this account, true to his respect for others’ privacy, Carl also protected the name of the letter’s author. Margulis subsequently shared her letter with Poundstone and Davidson.

It is only fair to tell the reader that Carl Sagan was my colleague and friend for more than a decade. Like all of us, he was flawed. In one of his last popular books, the grim *Demon-Haunted World* (Ballantine, 1997), Carl offered a list of his scientific mistakes. In the same book, he also reiterated one of his greatest concerns: “we humans have a talent for deceiving ourselves. Skepticism must be a component of the explorer’s toolkit, or we will lose our way.” Biographers might pay heed to this warning. ■

Christopher Chyba is at the SETI Institute, 2035 Landings Drive, Mountain View, California 94404, and in the Department of Geological and Environmental Sciences, Stanford University, Stanford, California 94305, USA.



Life on Mars: a Martian ‘petrophage’, inspired by Sagan and Joshua Lederberg’s assessments preceding the Viking mission to Mars.

Innovation machine on the boil

Basic research and technological application are moving closer together.

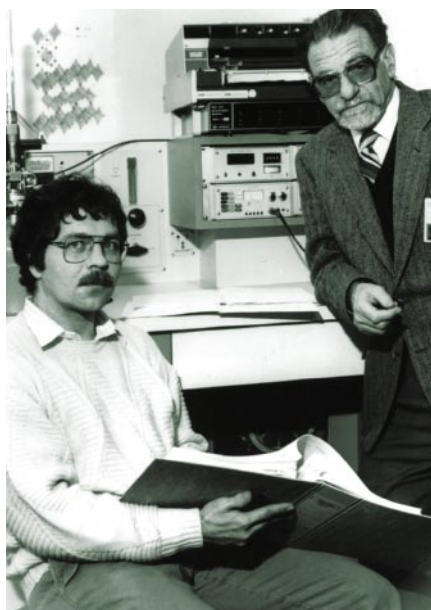
Helga Nowotny

One of the most exciting recent success stories of science began in September 1986 with the appearance in the *Zeitschrift für Physik* of an article with the cautious title "Possible high- T_c superconductivity in the Ba-La-Cu-O system". A few weeks later, the names of the two authors — Alexander Müller (pictured right) and Georg Bednorz (left) — and their discovery hit the front page of *The New York Times*. Researchers around the world were caught in an unprecedented frenzy, trying to replicate and surpass the findings of the initial breakthrough. The race for higher temperatures was on.

In early January 1987 Paul Chu found YBaCuO superconducting at about 90 K. He immediately applied for a patent on the new compound, which marked a technological breakthrough because the critical temperature could now be attained with liquid nitrogen. The euphoria peaked during the annual meeting of the American Physical Society in March 1987, when almost 3,000 physicists squeezed into the New York Hilton hotel to hear the latest news — an event that became known as the "Woodstock of physics".

Things began to settle somewhat, at least in scientific circles, when in October 1987 Müller and Bednorz were awarded the Nobel Prize in Physics for the discovery of the first ceramic material to superconduct at the then high temperature of 30 K. The award recognized the significance of their discovery, but it also meant that a stimulating race entered the next phase. It had become clear that years of more serious research still lay ahead before the new phenomenon would be fully understood. Surprisingly, less time would elapse before technologically reliable and commercially viable applications were to come. Views on how superconductivity is originated by electron-electron interactions, as opposed to electron-phonon interactions, are still unsettled. Theoretical understanding lags behind technological applications.

The discovery of high-temperature superconductivity (HTS) was unexpected in terms of its discoverers, the place of its discovery and the scientific ideas involved. It contradicted conventional wisdom and the expectations of peers and research administrators. Müller and Bednorz were outsiders; Müller an acknowledged specialist on perovskites, Bednorz a crystallographer. They benefited from the novice effect, but also enjoyed a degree of autonomy which, though temporary, allowed them to prepare for the unpredictable. Of IBM's three superconduc-



Müller and Bednorz were awarded a Nobel prize for the discovery of the first ceramic material to superconduct at 30 K.

tivity laboratories, Rüschlikon near Zürich, where the two researchers were based, was by far the most modestly equipped. And the discovery contradicted long-held views. It not only overturned the established empirical rules of Bernd Matthias, one of the grand old experimentalists in superconductivity, but also unveiled previously unknown phenomena not accounted for by the Bardeen-Cooper-Schrieffer theory.

But HTS also exemplifies the dramatic transformation that science is undergoing. The effects on the research system of the emergence of HTS can be likened to a building being tested by subjection to a transient load which reveals hidden strengths and weaknesses. The case of HTS shows how complex and fluid the situation has become. Researchers can no longer expect to find an environment unconditionally hospitable to their work. It takes an extraordinary amount of time and energy to set up the conditions under which research can run for a predictable period. Such efforts have become an

integral feature of the work of researchers. Furthermore, nowhere is science policy firmly in place. It is continuously shaped and reshaped under the impact of new and the constraints of old requirements, such as shifting priorities, strategic thinking on how to combine knowledge-driven with mission-oriented research, and how to accommodate objectives such as wealth creation.

HTS provides a fascinating glimpse into the present working of the scientific system as an 'innovation machine'. It reveals how basic research and technological applications are moving closer together and how the anticipation of technological potential affects the choice of what basic research is pursued. Basic research was once defined by psychologist Michael Posner as "that subset of what counts as good science which cannot be sold for immediate use". Yet HTS was 'sold' to the public — by the media, by politicians, by research-funding agencies and by the researchers themselves. All participants colluded in 'selling' the promise of technological advance as part of a widely shared belief. The case of HTS highlights the extent to which new criteria, extending beyond merely 'good science', have come into play and are effectively used in judging which areas of basic research should be funded.

The story of HTS also invites questions about the ideal conditions in which to foster individual creativity, and about how thin the line is between mere chance and programmed success. It probes the validity of Louis Pasteur's dictum that "chance favours only the prepared mind". It asks what preparedness means, not only on the individual level, but also on the institutional level. Even if we knew how to create conditions under which creativity can flourish, and how to favour the occurrence of what cannot be planned, the problem remains of how to turn highly individualistic bursts of scientific creativity into socially desired techno-scientific outcomes. For the most disturbing paradox is this: there has been a relative decline in the importance of the individual creative act, while its proliferation is encouraged. Individual scientific creativity has become a necessary, but no longer a sufficient, precondition in a long, branching sequence of possibilities. It has become contingent on what the innovation machinery of science will make of it as each new breakthrough is made. ■

Helga Nowotny is at the Swiss Federal Institute of Technology, ETH-Zentrum, CH-8092 Zürich, Switzerland. She and Ulrike Felt are the authors of *After the Breakthrough* (Cambridge Univ. Press, 1997).

Role reversal in geomagnetism

Bruce Buffett

What drives the erratic reversal of Earth's magnetic field? Simulations that allow for variations in heat flow at the core–mantle boundary now implicate events in the mantle.

A record of the Earth's magnetic field over geological time is preserved in rocks that are magnetized at the time of their formation. Changes in the direction and intensity of this fossil magnetism indicate that the Earth's magnetic field has repeatedly reversed its polarity. More remarkable is the variability in the frequency of reversals. In some instances the field was stable for tens of millions of years; in others it reversed erratically after only a few thousand years. On page 885 of this issue¹, Glatzmaier and colleagues show that the origin of this variability may lie outside the liquid core, where the Earth's magnetic field is thought to be generated and where the causes of reversals have hitherto been largely sought.

Gradual changes in the average reversal frequency are well documented². The record of reversal over the past 160 million years (Myr) reveals a steady decrease in reversal frequency prior to a prolonged stable interval, known as the Cretaceous Superchron (Fig. 1). Subsequent increases in reversal frequency have continued until the present day. Variations in reversal frequency over periods of 10^6 years have always been difficult to reconcile because timescales typically associated with generation of the magnetic field in the Earth's core are relatively short. For example, the time required for liquid iron to rise through the core and sustain the magnetic field is probably several hundred years. In the absence of this flow, the field would dissipate in 10^4 years. Timescales of 10^6 years are much more characteristic of convection in the overlying mantle, which drives plate tectonics at the Earth's surface. Ocean basins can open or close on this timescale and we also expect major rearrangements of temperature anomalies in the mantle. It is natural then to attribute long-period changes in the core to changes in the mantle, but the nature of the interaction has been difficult to quantify.

The principal difficulty arises from our poor understanding of the dynamics of the liquid core. Consequently, suggestions about the possible responses of the core to changing conditions in the mantle are quite speculative. However, advances in computer simulations that properly account for the interplay between convection and magnetic-field generation in the core now make it feasible to explore the role of the mantle with greater confidence³.

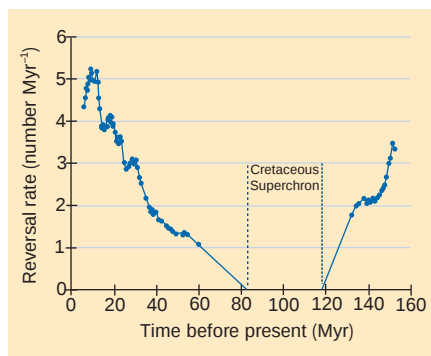


Figure 1 Record of average reversal rate over the past 160 million years (Myr) (ref. 2). Rates are averaged over a sliding interval of time to emphasize the long-period trend. A decrease in reversal rate precedes the period of prolonged stability between 118 and 83 Myr before present. Subsequent increases in reversal rate over the past 80 Myr now yield reversals every 200,000 years on average. Glatzmaier *et al.*¹ argue that such long-period variations in the behaviour of the field reflect the influence of changing conditions in the Earth's mantle, which in turn affect the behaviour of the geodynamo.

Glatzmaier *et al.*¹ chose to look at the influences of temperature anomalies in the mantle rather than other types of mechanical or electromagnetic interactions that have been proposed⁴. Temperature anomalies of 10^2 K in the mantle will almost certainly influence the core, where temperature anomalies are more likely to be 10^{-3} K because of the very low viscosity of the liquid iron core⁵. To a first approximation, the temperature at the core–mantle boundary should reflect the uniformity of the temperature in the core. As cold material in the mantle sinks towards the boundary, temperature gradients should steepen and the heat flow across the boundary should increase locally. Conversely, warm material at the base of the mantle should lower the heat flow.

Glatzmaier *et al.* incorporate the influence of the mantle into their simulations by imposing heat-flow conditions that vary over the surface of the core–mantle boundary. They investigate several simple patterns of heat flow and compare the results with a model where uniform heat flow is imposed. They also consider a model that uses our best estimate of the current distribution of temperature anomalies in the mantle, based indirectly on relationships between temper-

ature anomalies and lateral variations in elastic properties. The latter are detected through their influence on the propagation of seismic waves⁶.

Differences between the various simulations are quite striking. Significant changes in the reversal frequency, and the average field strength and its short-period variability, occur when the pattern of heat flow is changed. The authors offer a simple but physically appealing interpretation of their results. They note that variations in heat flow at the boundary induce a flow in the core, much like atmospheric circulation driven by differential solar heating of the Earth's surface. This flow is superimposed on and interacts with the deeper convective flow that sustains the magnetic field.

When the induced flow reinforces the underlying convective flow, the field strength increases and its variability decreases; the dipole component of the field remains more closely aligned with the geographic pole and reversals occur less frequently. The opposite occurs when the induced flow opposes and disrupts the convective circulation. In detail, the changes in behaviour must depend on nonlinear interactions between the two flows and the magnetic field, but the simulations clearly demonstrate that changes in heat flow conditions can strongly influence the behaviour of the magnetic field. In particular, the change in behaviour appears to be sufficient to explain the variations observed in Fig. 1.

Curiously, the simulation that produces a magnetic field most like that of the present-day field of the Earth has uniform heat flow at the boundary. Heat-flow conditions that are thought to be more realistic yield a magnetic field with too much temporal variability between reversals. Perhaps the assumed relationship between temperature and elastic properties overestimates the size of thermal anomalies in the mantle. Suggestions that lateral variations in elastic properties in the mantle are due partly to chemical variations⁷ support this view, but it is equally likely that the reported discrepancy indicates shortcomings in the numerical simulation. Despite the enormous computational effort required to carry out the simulations in this study, the range and scale of parameters accessible to numerical calculations is still far from those of the Earth. Different numerical strategies for approximating Earth-like

conditions yield magnetic fields with significantly different structures inside the core⁸, so there is good reason to be cautious when interpreting the results of the numerical simulations.

On the other hand, there are reasons for optimism. The simulations predict increased variability between reversals when the reversal frequency is high, consistent with observations. There is also evidence in the simulations for so-called excursions, where the field appears to reverse partially before jumping back to the initial polarity. Similar behaviour has been reported in the Earth's magnetic field⁹. Simply, the fact that we can compare numerical predictions with observations from past reversals opens new

lines of enquiry that were inconceivable just a few years ago.

Bruce Buffett is in the Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada. e-mail: buffett@eos.ubc.ca

1. Glatzmaier, G. A., Coe, R. S., Hongre, L. & Roberts, P. H. *Nature* **401**, 885–890 (1999).
2. Merrill, R. T., McElhinny, M. W. & McFadden, P. L. *The Magnetic Field of the Earth: Paleomagnetism, the Core, and the Deep Mantle* (Academic, San Diego, 1996).
3. Glatzmaier, G. A. & Roberts, P. H. *Nature* **401**, 203–208 (1995).
4. Hide, R. *Nature* **222**, 1055–1056 (1969).
5. Stevenson, D. J. *Geophys. J. R. Astron. Soc.* **88**, 311–319 (1987).
6. Su, W.-J. *et al.* *J. Geophys. Res.* **96**, 6945–6980 (1994).
7. Kellogg, L. H. *et al.* *Science* **283**, 1881–1884 (1999).
8. Kuang, W.-J. & Bloxham, J. *Nature* **389**, 371–374 (1997).
9. Gubbins, D. *Geophys. J. Int.* **137**, F1–F3 (1999).

Neurobiology

Exciting neurotrophins

Benedikt Berninger and Mu-ming Poo

The neurotrophins are a family of secreted proteins named after their ability to regulate the survival and differentiation of nerve cells ('neuron' means nerve and 'trophe' means nutrient). But there is also evidence¹ that neurotrophins are involved in the processes underlying neuronal plasticity — in particular, they cause rapid changes in both the shape and physiology of nerve cells². Neurotrophins can acutely modulate synaptic processes such as the secretion of neurotransmitter-filled vesicles from nerve endings. And whereas the survival and differentiation functions are usually evident over a period of hours or days, these synaptic effects are detectable within few minutes³.

A study by Kafitz *et al.*⁴ on page 918 of this issue shows that neurotrophins may also have effects that were previously assigned to classical excitatory neurotransmitters. They show that neurotrophins can cause membrane depolarization within a few milliseconds, eventually leading to the firing of an action potential. These new data imply that, at certain synapses, neurotrophins not only modulate chemical transmission but they are the very compounds that mediate it.

Over the past ten years, the idea that neurotrophins might regulate synaptic processes has gained momentum. First, the biosynthesis and secretion of neurotrophins such as brain-derived neurotrophic factor (BDNF) are enhanced on electrical or synaptic stimulation¹. Second, neurotrophins can regulate synaptic function within a few minutes, involving both pre- and postsynaptic modifications². Finally, mice that lack the *BDNF* gene have impaired long-term potentiation⁵ (a model for the cellular processes that underlie learning and

memory), consistent with the idea that BDNF is involved in synaptic plasticity.

Even so, the discovery that neurotrophins can cause membrane depolarization — turning them from mere modulators to potential mediators of synaptic transmission — is a surprise. Kafitz *et al.*⁴ took slices from the hippocampus, cortex and cerebellum of juvenile rats, then puffed minute amounts of BDNF onto the cell bodies of various types of neurons while recording their membrane potentials. The authors observed that, in less than ten milliseconds, BDNF caused a rapid depolarization which, depending on the dose, eventually resulted

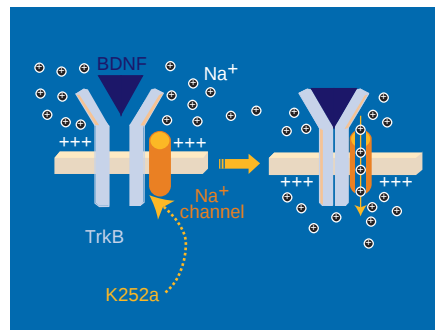


Figure 1 Model for membrane depolarization by brain-derived neurotrophic factor (BDNF), proposed by Kafitz *et al.*⁴. On binding of BDNF, the TrkB receptor interacts directly (or indirectly through an adaptor molecule) with the tetrodotoxin-insensitive sodium channel and causes it to open. The resulting sodium influx leads to membrane depolarization and, eventually, action-potential firing. According to this model, the Trk inhibitor K252a has to act by interfering with the receptor–channel interaction rather than by blocking phosphorylation-dependent steps.

in trains of action potentials being fired. The same type of response was seen when the authors applied puffs of glutamate, the main excitatory neurotransmitter at central synapses. However, BDNF elicited the effect at much lower concentrations than glutamate (in the order of three magnitudes), making BDNF the most potent natural neuroexcitant identified to date.

But how do neurotrophins cause membrane depolarization? Kafitz and colleagues showed that this type of response is mediated by tyrosine kinase receptors of the Trk family. The authors could block depolarization by prior treatment with K252a, a relatively selective inhibitor of Trk receptors. Furthermore, NT-4/5 was just as potent as BDNF in causing membrane depolarization. Both of these neurotrophins are ligands to TrkB, which is known to be expressed in the neurons examined. In contrast, nerve growth factor (a TrkA ligand) had no effect, consistent with the absence of TrkA in these cells.

So what signal follows TrkB activation to cause membrane depolarization? In voltage-clamp experiments, Kafitz *et al.* found that the current induced by BDNF reversed at the sodium equilibrium potential. This suggests that BDNF activates sodium channels via TrkB. The second messengers or phosphorylation-dependent steps that are usually thought to underlie signalling by TrkB probably do not mediate this response, because BDNF works extremely quickly and prolonged wash-out of the intracellular milieu does not cause run-down of the response. Instead, the authors propose that TrkB might interact with the channel protein — either directly, or indirectly through an adaptor molecule — and cause it to open. In that case, perhaps K252a blocks the response by steric hindrance, preventing TrkB from binding to either the channel or the adaptor (Fig. 1).

If the model of an interaction between the TrkB receptor and a sodium channel is correct, it is a new way of coupling transmembrane receptors to the membrane potential. It differs from ionotropic receptors, where the receptor complex also forms the channel, and from metabotropic receptors, which regulate membrane conductances through guanine-nucleotide-binding (G) proteins over a much slower timescale. The sodium channel that is activated by BDNF is tetrodotoxin-insensitive and is probably rapidly desensitized (because BDNF must be applied quickly to detect the effect). This could explain why the depolarizing action has escaped detection in previous electrophysiological studies.

But perhaps physical interaction between neurotrophin receptors and ion channels is more common than first thought. In last month's *Neuron*, Li *et al.*⁶ reported that BDNF can regulate ion-channel activity on a slower timescale than that of the sodium channel described by Kafitz *et al.*⁴. In neurons

from the pons, BDNF activates a cation channel known as TRPC3. As well as in its delayed onset, the BDNF-induced current through the TRPC3 channel differs in several aspects from that described by Kafitz and co-workers. For instance, it requires signalling through a second-messenger cascade. Intriguingly, TRPC3 can be co-immunoprecipitated with TrkB in protein extracts from neonatal rat brain, suggesting that TrkB might also bind, directly or indirectly, to this putative channel.

One of the most exciting aspects of Kafitz and colleagues' study is the possibility that certain synapses are 'neurotrophinergic' — that is, they use neurotrophins such as BDNF to mediate chemical synaptic transmission instead of (or as well as) more classical neurotransmitters. Where to look for such synapses? Although much is known about localization of the messenger RNAs that encode neurotrophins and their receptors, the location of the proteins themselves is more elusive. Are neurotrophins expressed within or outside the synapse? Are they released from the pre- or the postsynaptic neuron? What kinds of stimuli would lead to secretion of the right amounts of neurotrophins to depolarize the membrane? If released, do neurotrophins act in an autocrine or a paracrine manner? (An interesting system in which to start investigating this may be the synapse between the so-called mossy fibres and the CA3 pyramidal neurons in the hippocampus, because the mossy-fibre axons are thought to contain high levels of BDNF.)

It will also be important to assess whether the depolarization induced by BDNF can account for the rapid effects of neurotrophins on synaptic transmission that have been described previously^{2,3}. For instance, neurotrophins could increase the levels of presynaptic calcium by activating voltage-gated calcium channels through membrane depolarization. Such increased calcium could explain both the increased frequency of miniature synaptic events and the potentiation of evoked synaptic transmission that are often found after exposure to neurotrophins^{2,3}.

Finally, what could the use of neurotrophins as neurotransmitters mean, physiologically speaking, for a synapse? If neurotrophins were released from the postsynaptic neuron, they could depolarize the presynaptic membrane and allow for reciprocal transmission. Because the levels of BDNF depend on a neuron's activity, neurotrophin-mediated depolarization may come into play when neuronal activity is increased, and, in particular, under pathological conditions such as seizures. Whatever the answers to all of these questions, we can guarantee that the excitement brought about by neurotrophins will be sustained for some time to come. ■

Benedikt Berninger and Mu-ming Poo are in the Department of Biology, University of California San Diego, 9500 Gilman Drive, La Jolla, California 92093-0357, USA.
e-mail: mpoo@biomail.ucsd.edu

Mathematics

Holes and hot spots

Ian Stewart

In the rapid development of mathematical physics between 1750 and 1825, the most important single element was the Laplace operator $\Delta = \partial^2/\partial x^2 + \partial^2/\partial y^2$ and its three-dimensional analogue. Even now, study of this partial differential operator, devised by Pierre Simon Laplace (pictured), continues to throw up surprises, and the latest comes from Krzysztof Burdzy and Wendelin Werner in *Annals of Mathematics* (149, 309–317; 1999). They tackle the so-called 'hot spots' conjecture, a deceptively simple question that was first posed in 1974 by J. Rauch. This conjecture is important for theoretical understanding of the Laplace operator, but its applications are likely to be indirect.

The Laplace operator lies at the heart of classical theories of heat, light, sound, electricity, magnetism, gravitation and fluid mechanics. It led to a profound new understanding of wave phenomena and diffusion phenomena. In particular, Joseph Fourier's theory of heat, which is a diffusion phenomenon, centres upon properties of the Laplace operator. In the subsequent 250 years mathematicians have made intensive studies of the Laplace operator and its generalizations — even the fashionable theory of superstrings can profitably be viewed in such a setting.

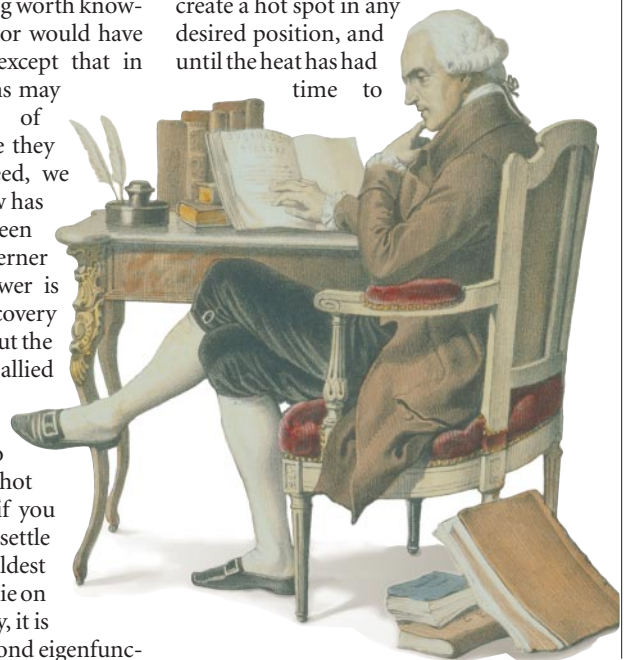
It might seem that anything worth knowing about the Laplace operator would have been discovered long ago, except that in mathematics simple questions may require the development of sophisticated methods before they can be answered. And indeed, we have much to learn. Only now has the hot spots conjecture been solved, for Burdzy and Werner show that the expected answer is wrong. The key to their discovery was some physical insight about the geometry of bicycle wheels, allied with an impressive command of technique.

Interpreted according to the theories of Fourier, the hot spots conjecture states that if you wait for initial transients to settle down, then the hottest and coldest regions of any planar domain lie on its boundary. More technically, it is about the geometry of the second eigenfunc-

1. Thoenen, H. *Science* **270**, 593–598 (1995).
2. McAllister, A. K. et al. *Annu. Rev. Neurosci.* **22**, 295–318 (1999).
3. Lohof, A. M., Ip, N. Y. & Poo, M.-m. *Nature* **363**, 350–353 (1993).
4. Kafitz, K. W. et al. *Nature* **401**, 918–921 (1999).
5. Korte, M. et al. *Proc. Natl Acad. Sci. USA* **92**, 8856–8860 (1995).
6. Li, H.-S., Xu, X.-Z. & Montell, C. *Neuron* **24**, 261–273 (1999).

tion of the Laplace operator. An eigenfunction of the Laplace operator is a function f such that Δf is a constant multiple of f . The constant here is called an eigenvalue. The patterns formed by a thin layer of sand on a vibrating metal plate — the celebrated 'Chladni plate' experiment found in most science museums — give a picture of such eigenfunctions, with the plate as the domain. Sand piles up along the 'nodal lines', where the eigenfunction vanishes.

A great deal is known about the eigenfunctions of the Laplace operator. For any domain, they are infinite in number, and their eigenvalues form a discrete sequence of numbers. In a musical analogy, they represent the fundamental and its higher harmonics — the basic components of any sound that the vibrating plate can produce. The hot spots conjecture states that the second eigenfunction of the Laplace operator (on a planar domain and subject to so-called Neumann boundary conditions — no flux at the boundary) attains its maximum and minimum values on the boundary of the domain. The second eigenfunction dominates the heat distribution once transients have died away. It is necessary to ignore transients because initial conditions can be chosen to create a hot spot in any desired position, and until the heat has had time to



JEAN-LOUP CHARMET/SCIENCE PHOTO LIBRARY

diffuse away from that position, it will remain a hot spot.

There are many precedents for such theorems in the geometry of the Laplace operator, whence the conjecture. In forthcoming work, David Jerison and Nikolai Nadirashvili prove that the hot spots conjecture is true for any convex domain. R. Bañuelos and Burdzy have proved it for certain long, thin domains, not necessarily convex ones (*J. Funct. Anal.*, in the press).

Nevertheless, Burdzy and Werner show that the full conjecture is false (they also say that Jerison and Nadirashvili have a different proof of this). Their basic idea is to consider a domain that resembles a bicycle wheel: a hub, at least three long thin spokes, and a tyre. Start with the hub hot and the tyre cold. Think about the flow of cold rather than that of heat. Cold reaches the hot hub only through the thin spokes so, as time passes, the hottest spot is pushed towards the centre of the hub.

So far, so good: the trick is to prove all this. Burdzy and Werner first define a domain with three identical spokes each about 20 times as

long as they are wide, a hub and a tyre. Unlike a real bicycle wheel, however, the spokes extend well outside the tyre — and the actual shape has to be chosen very carefully for the ideas of the proof to work. This domain is a complicated polygon with the symmetry of an equilateral triangle: three rotations and three reflections. The domain has three holes, each bounded by part of the hub, part of the tyre and parts of two spokes. The symmetry makes it possible to use known results about eigenfunctions in symmetric domains at crucial stages of the proof. However, the actual domain to which Burdzy and Werner's results apply differs slightly: a cut is made through the domain just described, to reduce the number of holes to two.

Whether the hot spots conjecture is true for domains with only one hole, or none, remains open. The geometry of the Laplace operator does not reveal its secrets lightly, but it undoubtedly remains a hot topic. ■

Ian Stewart is in the Department of Mathematics, University of Warwick, Coventry CV4 7AL, UK.
e-mail: ins@maths.warwick.ac.uk

Ecology

Scaling, energetics and diversity

Robert J. Whittaker

There are few established laws in ecology. Most ecologists have to contend with systems of bewildering complexity, in which it is hard to separate the wood from the trees. Scientific disciplines without a good framework of rules and laws are generally considered the poor relations of their ilk, and, as human impacts on our planet appear to have reached alarming levels, the need for reliable ecological theory is hard to overstate. The discovery and demonstration of new laws in ecology — such as the Universal Growth Law for plants described by Enquist *et al.*¹ on page 907 of this issue — is thus particularly noteworthy.

The existence of power relations (allometric relationships), whereby biological variables tend to increase or decrease in relation to body mass raised to the power of a quarter or three quarters, has been known for over half a century. But only recently has the realization of the ubiquity and generality of this relationship, which occurs at many levels of biological organization, elevated it to the status of a biological law². One of the best-known quarter-power scaling relationships is that between body mass and metabolic rate in animals, which it now appears holds over an extraordinary 27 orders of magnitude of size². Enquist and colleagues' work¹ is the latest step in the determination of analogous allometric relationships for plants, with potentially wide-

-ranging implications for ecological science.

Enquist and colleagues' general case is that production within species declines in a uniform fashion with increasing size. Why this should be so, theoretically, can be explained in relation to the energy and materials essential to all biological processes. In previous analyses of allometric relationships in plants, reported in *Nature* earlier this year³, this team has shown how plants conform to a suite of scaling rules, and constitute distribution networks with a fractal-like (that is, self-similar) architecture. Physical and biological constraints on the distribution of resources (water and nutrients in solution) through these networks underlie the '3/4 power scaling' of metabolic rate with mass (metabolic rate = mass^{0.75}). Put another way, biological rates are ultimately limited by the rates at which energy and materials can be distributed between the surfaces where they are exchanged and the tissues where they are used.

Enquist *et al.* provide a demonstration of their theoretical model drawn from a sample of 2,283 trees of 45 species, measured at two points in time, 20 years apart, from seasonally dry tropical forest in Costa Rica. This shows that a large array of different species show the expected relationship of diminishing growth with increasing size. The biology and ecology of trees is, of course, highly variable — there are light-demanding and



100 YEARS AGO

The long-standing question as to the admittance of women into full fellowship of scientific societies was brought before a meeting of the Lady Warwick Agricultural Association for Women on Thursday last, and the following resolution, supported by a paper by Mrs. Farquharson, was adopted: "That it is desirable and important that duly qualified women should have the advantages of full fellowship in scientific and other learned societies, *e.g.* the Royal, the Linnean and the Royal Microscopical." The arguments in favour of and in opposition to this proposal have been stated so many times that most members of scientific societies are familiar with them. Six years ago the council of the Royal Geographical Society elected several ladies as fellows, but their action was disapproved at two special meetings, and resolutions to the effect that it was inexpedient to admit ladies as ordinary fellows were carried by conclusive votes. ... The time may of course come when women will be equally eligible with men for membership of the learned societies, but facts such as those cited show that there is distinct opposition to the admittance of women at present, and no sudden change of feeling can be expected, though individual cases of "duly qualified" women might be considered.

From *Nature* 26 October 1899.

50 YEARS AGO

All through the nineteenth century and well into the twentieth, mathematicians studied the problems suggested by the generation and propagation of waves on the surface of the sea and other bodies of water. But during all this time there was very little scientific observation of waves, and practically none of it was by means of specially designed measuring instruments. Oceanographers paid very little attention to the details of the phenomena. But during the Second World War, it was part of the preparation for the landing of troops and material on beaches that forecasts should be made of the state of swell and surf on those beaches at particular times... C.C. Bates, of the United States Navy, gives an account of the utilization of wave forecasting in the invasions of Normandy, Burma and Japan. Excluding cases in which forecasts of winds were in error, it was found that forecasts of waves on Normandy landing beaches were correct for 88 per cent of the time.

From *Nature* 29 October 1949.

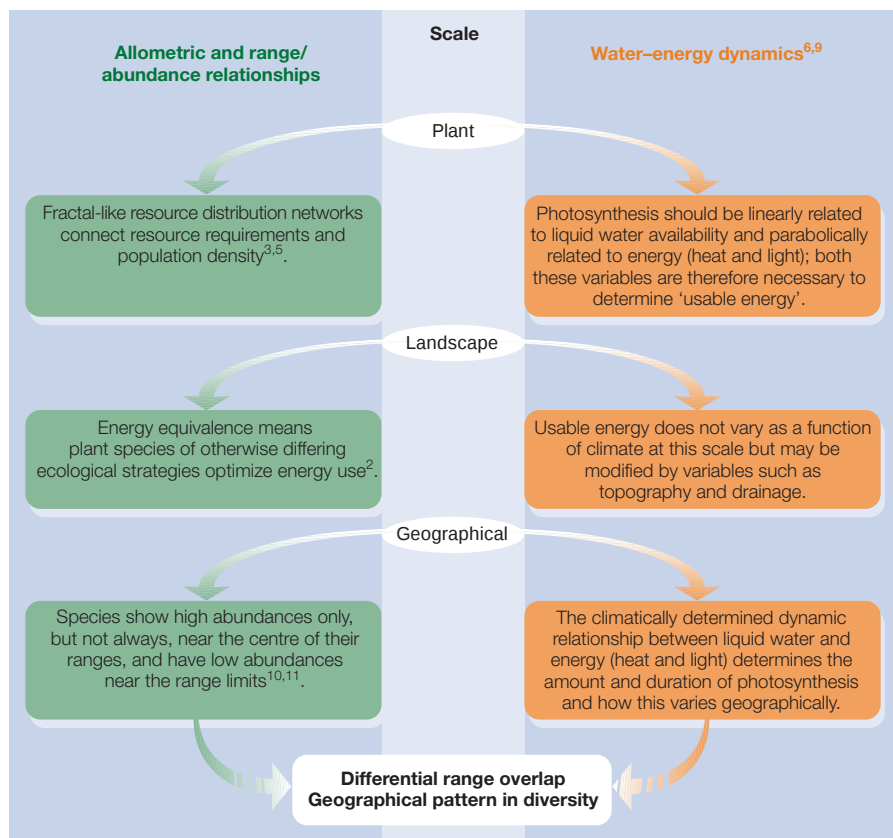


Figure 1 Macroecological insights into plant diversity. Differential range overlap produces geographical pattern (for example, latitudinal gradients) in diversity, wherein the species capable of persisting through time in a region are those able to optimize energy use¹. Climate determines the capacity for richness, consistent with richness at the macroscale being monotonically related to increasing usable energy (that is, productivity) and parabolically related to climatic energy⁶.

shade-tolerant species, insect- and animal-pollinated species, species adapted to differing soil conditions and so forth. In tropical forests, these and other traits occur in many different combinations, such that reduction of variation onto a single successional axis — pioneer to 'climax' — has proven impractical⁴. Similarly, no satisfactory general theory to explain variation in diversity has emerged from the study of these traits and related processes.

Enquist *et al.*¹ have shown, however, that when variation in wood density (itself an important trait reflecting differences in the allocation of metabolic energy) is taken into account, the data fit the theoretical power relation demanded by their model remarkably well. What this means is that usable energy is being deployed in a variety of growth-form, life-history and wood-density strategies that each amount to varying ways of attaining the same optimal use of energy. In other words, there is energy equivalence across species (see also ref. 5).

Within this seeming evolutionary constraint, there are interspecific differences in energy allocation over the lifespan. These take the form of differing rates of volume increment, differing lifespan and varying reproductive strategies, which together underpin the diversity of these tropical forests. The key

point in relation to diversity is that, in the large block of forest studied, there is clear evidence that energetics are crucial to the persistence of the diversity of species present, otherwise their growth in biomass would not exhibit such a consistent scaling relationship. Hence systems with more usable energy should allow for a greater subdivision of energy into more distinct plant niches. This ties in with another fundamental relationship, recently reported by E. M. O'Brien in the *Journal of Biogeography*⁶ — a capacity rule describing the climatic potential for richness (CPR). Both sets of

findings may be useful in developing a theory of diversity which, like the diversity problem itself, is trans-scalar in nature (Fig. 1).

In practice, although Enquist and colleagues' scaling relationships are detectable within trees and in smaller plant units, even down to the cell², the significance of energetics in regulating diversity at the local scale (for example, 0.1 hectare) is often undetectable, being obscured by other variables whose measurable heterogeneity occurs at local to landscape scales. The significance of variation in usable energy becomes evident, however, at the macro-scale, at which O'Brien's Interim General Model⁶ describes the CPR as a linear function of liquid water and a parabolic function of energy. Liquid water is also fundamental, not just as a vital ingredient in the process of photosynthesis, but also to the development of allometric models, as energy and nutrients are moved into and through plants in solution³.

So these scaling laws seem to provide a basis for the development of ecological theories that are trans-scalar in geographical space, something with which traditional branches of ecology have struggled. In separating the wood from the trees — at least in terms of correcting for density variations — Enquist *et al.* have shown the vitality of the emerging field of macroecology^{2,7,8}.

Robert J. Whittaker is in the School of Geography, University of Oxford, Mansfield Road, Oxford OX1 3TB, UK.

e-mail: Robert.Whittaker@geog.ox.ac.uk

- Enquist, B. J., West, G. B., Charnov, E. L. & Brown, J. H. *Nature* **401**, 907–911 (1999).
- Brown, J. H. & West, G. B. (eds) *Scaling in Biology* (Oxford Univ. Press, 1999).
- West, G. B. *et al.* *Nature* **400**, 664–667 (1999).
- Brown, N. D. & Jennings, S. in *Dynamics of Tropical Communities* (eds Newbery, D. M., Prins, H. T. & Brown, N. D.) 79–94 (Blackwell Science, Oxford, 1998).
- Damuth, J. D. *Nature* **395**, 115–116 (1998).
- O'Brien, E. M. *J. Biogeogr.* **25**, 379–398 (1998).
- Brown, J. H. *Macroecology* (Univ. Chicago Press, 1995).
- Maurer, B. A. *Untangling Ecological Complexity: The Macroscopic Perspective*. (Univ. Chicago Press, 1999).
- O'Brien, E. M. *J. Biogeogr.* **20**, 181–198 (1993).
- Stevens, G. C. & Enquist, B. J. in *Ecology and Biogeography of Pinus* (ed. D. M. Richardson) 183–190 (Cambridge Univ. Press, 1998).
- Enquist, B. J. *et al.* *Evol. Ecol.* **9**, 586–604 (1995).

Mutation

Enzymes of evolutionary change

Miroslav Radman

The future is unpredictable. Survival therefore depends on the generation of a large repertoire of biological diversity — the evolutionary equivalent of buying a large number of lottery tickets. Natural selection can be compared with drawing the lottery, in that a strong selection event reduces diversity because it allows only a small fraction of the original population to survive (the winners).

Diversity then needs to be restored to allow future selections to be faced. In other words, genetic diversification needs to be a continuous process, and reports in *Proceedings of the National Academy of Sciences*¹, *Molecular Cell*² and the *Journal of Biological Chemistry*³ show how this might be done.

Genetic diversity is generated by mutagenesis, which causes changes in the DNA

sequence. A classic example of the generation of genetic diversity is the interaction between infectious agents and the immune system. Viruses and bacteria mutate extensively in an attempt to generate rare variants that can escape the host's immune system. In turn, the immune system mutates to try to create antibodies that recognize these new variants. Similarly, presumptive tumour cells must accumulate mutations in several genes in order to evade the systems that monitor and block unrestrained cell growth. Clearly, these and similar adaptive processes are limited by the supply of mutations (Fig. 1).

Most newly arising mutations either have no apparent effect or they are harmful. So mutagenesis has traditionally been viewed as an unavoidable consequence of imperfections in the processes of DNA replication and repair. But if diversity is essential to survival, and if mutagenesis is required to generate such diversity, perhaps mutagenesis has been positively selected for throughout evolution. The pioneering work of Nelson and colleagues^{4,5}, along with the new papers¹⁻³, supports this idea with the identification of several enzymes designed to generate mutations.

These enzymes — let's call them DNA mutases — all belong to a special group of DNA-synthesizing enzymes, the DNA polymerases. Unlike the replicative DNA polymerases, which faithfully copy DNA sequences, mutases produce errors at high rates. I predicted the existence of such mutases more than a quarter of a century ago⁶, as a key element of the SOS system. This inducible, genetically programmed process allows individual cells to mutate when their survival is threatened, thereby increasing the genetic diversity and adaptability (fitness) of the endangered population⁶. But Evelyn M. Witkin was first to imagine that messy polymerases could be mutagenic⁷.

There seem to be several varieties of DNA mutase, including the *Escherichia coli* UmuCD' (DNA polymerase V)^{1,3} and DinB (DNA polymerase IV)² proteins, and the yeast REV1 (ref. 4), REV3/7 (DNA polymerase ζ)⁵ and RAD30 (DNA polymerase η)⁸. Bacterial DNA polymerase V (and the eukaryotic DNA polymerases ζ and η) can copy damaged DNA, allowing it to be replicated and the cell to survive. But DNA polymerase IV presumably acts on undamaged DNA, producing apparently 'gratuitous' mutations (Fig. 2). Both enzymes are inducible components of the SOS system.

All of these mutases are part of stress-inducible processes which allow them to function only when high mutation rates are advantageous⁶. This minimizes their genetic cost relative to 'constitutive' mutators — that is, mutants with defects in DNA repair or biosynthesis, or in error-correction by mismatch repair⁹. A hypothetical mechanism for limiting the mutagenic activity of DNA

polymerase V to damaged sites in DNA is shown in Fig. 2.

This mechanism involves the formation of a nucleoprotein filament between single-stranded DNA and the recombination protein RecA*. The work of Tang *et al.*¹ and Bacher Reuven *et al.*³ indicates that such a nucleoprotein filament may be a poor substrate for elongation by the real chromosomal replicase — the DNA polymerase III holoenzyme. Moreover, the RecA* filament seems to be the sole substrate for the DNA polymerase V mutase; only in the presence of RecA* can this polymerase copy a damaged DNA template^{1,3}. But once the gap has been filled in, the RecA* filament vanishes and the new DNA chain is again elongated by DNA polymerase III.

The role of DNA polymerase IV (DinB) in mutagenesis is less clear. The only known effect of *dinB* mutants is a reduction in viral mutagenesis, normally observed when a virus infects a bacterial host in which the SOS response has been induced ('untargeted mutagenesis')¹⁰. Because untargeted mutations appear initially as genuine mispaired or unpaired bases^{2,11}, DNA polymerase IV is the leading candidate for a genuine, genome-wide mutase acting under a variety of meta-

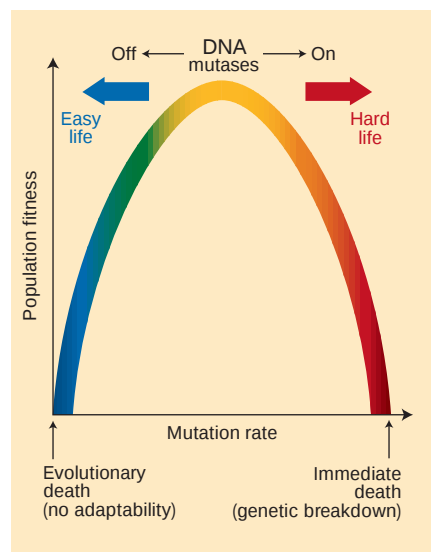


Figure 1 Mutation rates and genetic adaptability (fitness). Although the optimal mutation rate for a perfectly adapted clonal population is close to zero, with no mutations the population could not adapt to environmental changes. Conversely, a high mutation rate is optimal for a population under strong selective pressure, but too many mutations would cause a genetic breakdown. In a bacterial population, the global mutation rate can be influenced by the variable fraction of genetic mutators, which are individually in the red zone¹³. The newly identified mutases¹⁻⁵, present in all cells, produce mutations only when a genetic or metabolic stress triggers their induction and activation. The rare, specific mutations that they create, which save the organism in times of stress, are adaptive mutations.

bolic stresses^{9,12,13}.

A common characteristic of both *E. coli* mutases is that they are inefficient DNA polymerases both *in vitro*¹⁻³ and *in vivo* (where they require the help of other polymerases)^{14,15}. The precision of the replicative DNA polymerases results largely from their inability to extend mismatched DNA ends — once they have made a mistake they abandon the DNA. These mismatched ends then become substrates for a polymerase-associated 'proofreading' activity, which removes the misincorporated nucleotide and allows the polymerase to continue extending the DNA. But in times of stress, or at sites of DNA damage, these termini may be extended by the DNA mutases. The mutases fix the mutation and recreate a good substrate for the replicative polymerase^{2,14}. In other words, mutases may cause mutations not by incorporating the wrong bases, but by rescuing molecules that already contain these misincorporated bases.

Inducible transient mutator activity goes a long way towards reducing the genetic cost

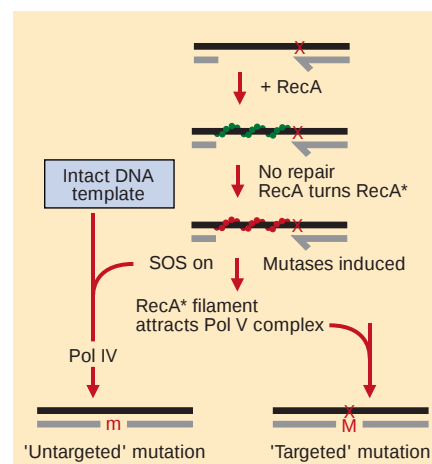


Figure 2 Possible mechanisms for controlling mutase activity in *Escherichia coli*. Persisting single-stranded DNA (here caused by damage in the DNA template) is rapidly covered by RecA (green), which searches for similar sequences to patch up the gap by recombination. When this cannot occur, the RecA filament changes and becomes a co-protease (RecA*; red), which activates the SOS system and induces the expression of over 20 life-saving genes. These include *umuC* and *umuD* (which encode the DNA polymerase V mutase) and *dinB* (which encodes the DNA polymerase IV mutase). The activated DNA polymerase V complex (UmuCD₂) can copy non-coding damage in the template only in the presence of RecA*, producing mutations (M) that are 'targeted' to the damaged region. But, once induced, DNA polymerase IV does not require RecA*¹⁰, so it can act on apparently intact DNA to produce 'untargeted' mutations (m). This scheme is supported by the observed correction of untargeted (but not of targeted) mutations by the mismatch-repair system¹¹.

of a dangerous process like mutagenesis. But this cost could be further reduced if mutases could be targeted to genes under selective pressure — a bit of Lamarckism in the midst of Darwinian evolution. An example of such targeting might be the 'contingency genes' in some parasitic bacteria and protozoa. These genes encode surface antigens that are recognized by a host's immune system. Some of these antigens contain simple, repeated sequences (microsatellites), which are preferential sites for polymerase 'slippage' mistakes. In this way, the contingency genes have a higher mutation rate than surrounding sequences¹⁶ and can keep one step ahead of the immune system. DNA polymerase IV mainly produces such slippage mutations².

The evolutionary advantage of inducible mutases in microorganisms seems clear. But the genomes of complex eukaryotes — including humans — also contain sequences that resemble the DNA polymerase IV and DNA polymerase V mutases¹⁷. If such genes are active and cause mutagenesis, the somatic (non-reproductive) cells of, say, mice knocked out for such genes will be more sensitive than normal cells to the killing effects of radiation and chemical carcinogens because they will lack the mutagenic polymerases to replicate past DNA damage. But these cells will be resistant to the mutagenic and carcinogenic effects of DNA damage, because they will die rather than mutate. If this is the case, maybe people exposed to DNA-damaging radiation and chemicals (by, for example, cancer therapies, accidents and wars) could one day be protected by inhibiting mutase activity. Alternatively, if some mammalian mutases act in a targeted fashion only, or mainly, on immunoglobulin genes, they could be the long-sought key to generating the immense diversity of antibodies. ■

Miroslav Radman is at the Faculté de Médecine Necker-Enfants Malades, Université R. Descartes, 156 rue de Vaugirard, 75730 Paris Cedex 15, France.
e-mail: radman@necker.fr

1. Tang, M. *et al.* *Proc. Natl Acad. Sci. USA* **96**, 8919–8924 (1999).
2. Wagner, J. *et al.* *Mol. Cell* **4**, 281–287 (1999).
3. Bacher Reuven, N. *et al.* *J. Biol. Chem.* **274**, 31763–31766 (1999).
4. Nelson, J. R., Lawrence, C. & Hinkle, D. C. *Nature* **382**, 729–731 (1996).
5. Nelson, J. R., Lawrence, C. & Hinkle, D. C. *Science* **272**, 1646–1649 (1996).
6. Radman, M. in *Molecular and Environmental Aspects of Mutagenesis* (eds Prakash, L. *et al.*) 128–142 (C. C. Thomas, Springfield, Illinois, 1974).
7. Witkin, E. M. *Proc. Brookhaven Symp. Biol.* **20**, 17–55 (1967).
8. Johnson, R. E., Prakash, S. & Prakash, L. *Science* **283**, 1001–1004 (1999).
9. Taddei, F. *et al.* *Nature* **387**, 700–702 (1995).
10. Brotkor-Lannoye, A. & Maenhaut-Michel, G. *Proc. Natl Acad. Sci. USA* **83**, 3904–3908 (1986).
11. Caillet-Fauquet, P., Maenhaut-Michel, G. & Radman, M. *EMBO J.* **3**, 707–712 (1984).
12. Taddei, F. *et al.* *Nature* **387**, 700–702 (1997).
13. Cairns, J. & Foster, P. L. *Genetics* **128**, 695–701 (1991).
14. Bridges, B. A. & Bates, H. *Mutagenesis* **5**, 35–38 (1990).
15. Maenhaut-Michel, G. & Caillet-Fauquet, P. *Mutat. Res.* **230**, 241–254 (1990).
16. Moxon, R. E. & Wills, C. *Sci. Am. Jan.* 94–99 (1999).
17. McDonald, J. P. *et al.* *Genomics* **60**, 20–30 (1999).

Quantum mechanics

Putting an electron to rest

P. Meystre

The first cyclotron — developed by E. O. Lawrence at Berkeley in the 1930s — was designed to accelerate charged particles as they orbit with a circular motion in a magnetic field. Large cyclotrons built according to Lawrence's design became the first useful particle accelerators. Less familiar is the use of small, single-electron cyclotron accelerators for high-precision measurements of certain fundamental constants of nature. For example, measurements of the electron magnetic moment and the fine structure constant have provided some of the most accurate tests of quantum electrodynamics (QED) — the modern theory of how electrons and photons interact. In an experiment reported in *Physical Review Letters*¹, S. Peil and G. Gabrielse of Harvard University study the fundamental quantum limit of the cyclotron motion of a single electron successfully isolated from its surroundings. The ability to control this cyclotron motion offers increased accuracy in measuring the electron magnetic moment.

Uhlenbeck and Goudsmit first proposed in 1925 that the electron has an intrinsic angular momentum, or spin, with an associated magnetic moment. In Dirac's relativistic theory of the electron, this magnetic moment is given by $g\mu_B s$, where s is the electron spin with possible values of $\pm 1/2$ and g is the so-called gyromagnetic ratio precisely equal to two. Here $\mu_B = e\hbar/2m$ is the Bohr magneton, with $\hbar$ being the rationalized Planck's constant, m the electron mass and e

the fundamental charge. But Dirac's theory ignored the fact that a moving electron emits an electromagnetic field that feeds back on its motion. QED predicts that as a result of this back-action g is actually slightly larger than two. Accurate calculations and measurements of $g-2$ provide an extremely stringent test of QED, resulting in it being the best-tested theory in physics.

Precision measurements using trapped electrons have a long history², leading to more and more accurate results for $g-2$. But a few years ago, experimental progress was interrupted by the recognition that the electromagnetic vacuum surrounding the trapped electron was substantially different from its value in free space, for which g was computed. This resulted in uncontrolled cavity shifts that could not be accounted for by theory. To avoid this problem, Gabrielse and colleagues developed a new type of electron cavity, a cylindrical Penning trap, which is an almost ideal cylindrical resonator with well-understood electromagnetic modes³. In particular, these traps significantly reduce energy lost from the electron by spontaneous emission of radiation^{4,5}. The trap used by Peil and Gabrielse¹ reduces spontaneous emission by a factor of 140. This feat, together with the cooling of the cyclotron to eliminate excitation by background photons, promises a new generation of precise measurements of $g-2$.

In the cylindrical Penning trap the cyclotron motion of the electron is, to an excellent degree of approximation, harmon-

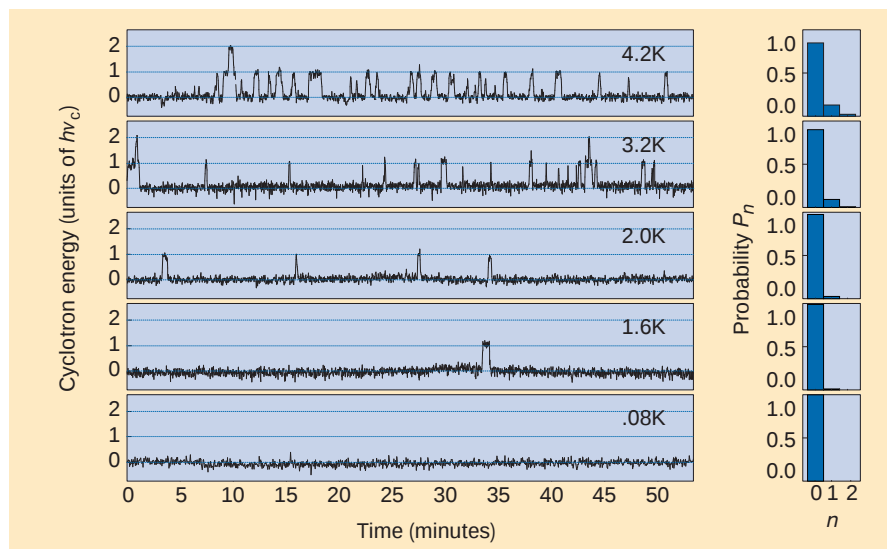


Figure 1 Time record of the quantum jumps between the lowest-energy quantized states of motion of a trapped electron observed by Peil and Gabrielse¹. As the temperature of the Penning trap is increased, so is the frequency of jumps. Whereas at low temperatures only jumps between the Fock states $|0\rangle$ and $|1\rangle$ are observed, jumps to higher-energy states become apparent at higher temperatures.

ic. That is, its oscillations are characterized by a single frequency, ω . As such it is an experimental realization of a simple quantized harmonic oscillator. This system plays a central role in physics and is very familiar to all students of quantum mechanics. It has an infinite number of evenly spaced energy levels labelled $|n\rangle$, where states numbered 0, 1, 2, ... are called Fock states, with the energy of the n th level equal to $E_n = (n + 1/2)\hbar\omega$. The ground state of the oscillator, $n = 0$, is characterized by the 'zero-point' energy, $\hbar\omega/2$.

In the Harvard experiment the electron cyclotron motion was cooled to a temperature of 80 mK. Although the internal structure of atoms makes it possible to cool them to even lower temperatures, this is, according to the authors, 50 times lower than previously achieved for an isolated elementary particle. At this temperature, it became possible to maintain the electron in its ground state for many hours at a time. The average time between transitions at 80 mK was estimated to be of the order of 10^{32} years! To observe quantum jumps between adjacent Fock states of cyclotron motion, it was necessary either to deliberately introduce resonant photons inside the cavity or to raise its temperature, which increases the background blackbody radiation (Fig. 1). Non-destructive detection of the cyclotron energy requires using a quantum non-demolition (QND) technique⁶, which means that repeated measurements will give the same answer. This is not generally true for observations of quantum systems because the back-action of the measuring apparatus affects their subsequent dynamics. The QND measurement here relied on a weak coupling imposed between the circular cyclotron motion of the electron and its axial motion along the central axis of symmetry of the trap, allowing the cyclotron motion to be observed indirectly.

Despite its theoretical simplicity, it has been surprisingly difficult to reproduce the simple harmonic oscillator in experiments: it requires a nearly perfect elimination of all sources of noise and perturbations. This can be achieved only by isolating the oscillator from its environment, a challenging task indeed. Nonetheless, several groups have nearly achieved a simple quantum harmonic oscillator, both in cooled trapped ions⁷ and in microwave photons trapped in resonators with very low loss rates^{8,9}. Although the ion experiments carried out by Wineland's group⁷ allow for a precise manipulation of the quantized centre-of-mass motion of the trapped ion, the detection of this motion has so far been destructive. This is in contrast to the microwave cavity QED experiments of Haroche and co-workers⁸, which are non-destructive and allow for repeated measurements of a single cavity photon. However, these experiments have so far involved at

most a one-photon state ($n = 1$). Experiments by Walther and colleagues⁹ use a micromaser configuration to generate higher photon states, which can in principle be detected by a QND technique. But the high energy dissipation in the micromaser has so far limited the purity of the states produced by this method.

The generation and manipulation of long-lived Fock states of the harmonic oscillator offer many fascinating directions of research, from quantum information processing, such as the realization of quantum logical gates, to studies of quantum decoherence and measurement theory as well as high-precision measurements. In particular, Peil and Gabrielse are already working on a more accurate measurement of $g-2$, but

more can be expected from the 'quantum-jump spectroscopy' made possible by these experiments. ■

P. Meystre is at the Optical Sciences Center, University of Arizona, Tucson, Arizona 85721, USA. e-mail: pierre.meystre@optics.arizona.edu

1. Peil, S. & Gabrielse, G. *Phys. Rev. Lett.* **83**, 1287–1290 (1999).
2. Brown, L. S. & Gabrielse, G. *Rev. Mod. Phys.* **58**, 233 (1986).
3. Gabrielse, G. & Tan, J. N. in *Cavity Quantum Electrodynamics* (ed. Berman, P. R.) 267 (Academic, Boston, 1994).
4. Haroche, S. in *Fundamental Systems in Quantum Optics* (eds Dalibard, J., Raimond, J. M. & Zinn-Justin, J.) 767 (North-Holland, Amsterdam, 1992).
5. Meystre, P. in *Progress in Optics* vol. 30 (ed. Wolf, E.) (Elsevier, Amsterdam, 1992).
6. Braginsky, V. B., Vorontsov, Y. & Thorne, K. S. *Science* **209**, 547–557 (1980).
7. Meekhof, D. et al. *Phys. Rev. Lett.* **76**, 1796–1799 (1996).
8. Nogues, G. et al. *Nature* **466**, 239–242 (1999).
9. Weidinger, M., Varcoe, B. T. H., Heerlein, R. & Walther, H. *Phys. Rev. Lett.* **82**, 3795–3798 (1999).

Immunogenetics

Soaring costs in defence

Peter Parham

For evolution of biological function, the processes of genetic recombination that cut, splice, move and replace segments of DNA, can be speedier agents of change than point mutation. Under pressure from diverse and cunning pathogens, the immune system exploits recombination to make and modify the families of receptors it uses to recognize antigens. Most heavily studied of the immune system's gene families are the class I and II genes of the human major histocompatibility complex (MHC) which influence donor choice in organ transplantation and susceptibility to chronic disease¹. These highly polymorphic genes were first studied as white-cell antigens of the blood which led to the HLA name, standing for human leukocyte antigen, now given to the whole human MHC. For more than 30 years immunogeneticists have construed maps of the HLA complex with ever-increasing elaboration. Now a paper² on page 921 of this issue by 'The MHC sequencing consortium' gives them the real thing: a complete nucleotide sequence of an HLA complex and a linear map ordering all of the genes.

Polymorphic MHC class I and II glycoproteins bind peptides derived from a pathogen's proteins and with them tell T cells to weigh in on the intruder. During disease epidemics some forms of class I and II MHC molecule stimulate T-cell responses that better favour survival, but which ones they are depends upon the particular infection. Consequently, human populations that were geographically separated and have different disease histories, differ in the number, sequences and frequencies of HLA class I and II alleles.

Over the much longer periods of evolu-

tionary time that separate humans from other vertebrate species, the idiosyncratic effects of selection by pathogens become more exaggerated. In mammalian species diverged by tens of millions of years the MHC differences extend beyond modification of alleles to alteration in gene number, structure and order within the MHC³. Exploring what happens during even longer divorce is a paper in this issue (page 923) by Kaufman et al.⁴, who determined the nucleotide sequence of a chicken MHC and compared it to the human one. In the 300 million years since chickens and colonels shared common ancestors, their MHCs have evolved along wildly different trajectories: the former emphasizing economy, efficiency and teamwork in their genes; the latter going for expansion, extravagance and individualism.

Birds are renowned for their frugality with DNA⁵ and in this the chicken MHC is no exception. While the HLA complex comprises 3,600 kilobases embracing 128 functional genes and 96 pseudogenes, the chicken MHC (called the B locus) has only 19 genes in 92 kilobases. Everything to do with the chicken MHC bespeaks of getting the biggest bang for the cluck: introns and intergenic regions are short, pseudogenes absent and repetitive elements few. Humans and chickens also differ in how they organize the essential MHC components. In the HLA complex the region containing class I genes is separated from the region containing class II genes by 800 kilobases of the class III region. By contrast in the chicken MHC the four genes of the class I region snuggle up to the seven genes of the class II region while the two class III region genes are on the outside, flanking the class I region. A further

point of discord: whereas the chicken genes for TAP, the supplier of class I-binding peptides, are in the class I region, the human ones lie in the class II region^{2,4}.

The vast sprawl of the human MHC means meiotic recombination between polymorphic genes occurs at rates of up to 2%. This mechanism generates new HLA haplotypes and is the source of the enormous variety of HLA phenotypes in the human population. By contrast the compact size of the chicken MHC essentially eliminates meiotic recombination as a force for generating MHC diversity, explaining why previous pedigree analysis had failed to find a single recombinant MHC haplotype⁶. As emphasized by Kaufman *et al.*⁴, absence of recombination raises the probability that alleles for different genes within the same haplotype have co-evolved to cooperate in function. One way this could be manifest is for TAP allotypes to specialize in the delivery of peptides that bind well to the linked class I allotypes. On the downside, low-frequency recombination and small numbers of class I and II genes are properties that make the chicken MHC less plastic and adaptable to environmental change than its human counterpart. These features could therefore explain why associations of MHC haplotype with resistance and susceptibility to infectious disease are much stronger in inbred chickens than other laboratory animals or humans⁷.

Some 40% of the expressed genes in the human MHC work for the immune system². Although this level of commitment is unlikely to be a uniform feature of the human genome, other parts of it are clearly conscripted to defence. Well established at centre stage, along with the MHC, are the large arrays of rearranging gene segments that determine B-cell immunoglobulins and T-cell receptors. Now emerging from the wings are two families of conventional, non-rearranging genes that provide receptors for natural killer (NK) cells. These large lymphocytes provide defence from the start of a pathogen's attack, a response distinguishing them from B and T cells which only become useful after days of infection.

One form of NK-cell receptor resembles a C-type lectin and is specified by gene families in the natural killer complex (NKC) on human chromosome 12 (ref. 8). A second form of NK-cell receptor is constructed from immunoglobulin-like domains and is specified by gene families in the leukocyte receptor complex (LRC) on human chromosome 19 (ref. 9). Whereas some gene families in the LRC exhibit haplotypic and allelic polymorphism¹⁰ those of the NKC are relatively conserved. An NK cell expresses several receptors drawn from these families and diversity within a person's NK-cell population arises from the expression of different receptor combinations. Some NK-cell

receptors bind polymorphic determinants of HLA class I molecules and appear sensitive to the effects that infections have upon them. Because the MHC, NKC and LRC are on different chromosomes their polymorphisms segregate independently in the human population. Consequently some people express NK-cell receptors for which they have no MHC class I ligand, while others have the ligand but not the receptor. The drawback to this arrangement is allayed by having NK cells express several different receptors, so at least one of them binds a person's MHC class I.

Although human genes for NK-cell receptors and MHC class I ligands are strictly partitioned between different chromosomes, that is not so in the chicken. Kaufman *et al.*⁴ describe two genes in the chicken MHC which specify proteins resembling the lectin-like receptors of mammalian NK cells. Such juxtaposition could also have fostered co-evolution producing NK-cell receptors that preferentially interact with class I allotypes encoded by the same MHC haplotype. Kaufman *et al.* take this idea one step further in their speculation that polymorphisms affecting NK-cell function are at the heart of the chicken's MHC-determined susceptibility to infection by the herpes virus that causes Marek's disease⁷. The idea is not so wild because an NK-cell-deficient patient presented with life-threatening herpes virus infections¹¹, and susceptibility to mouse cytomegalovirus, another type of herpes virus, maps to the NKC⁸.

The two papers in this issue^{2,4} vividly illustrate that working MHCs come in all shapes and sizes. The HLA complex is the sort of muddled mix of chaotic junk and organized function that one has come to expect from diverse and changeable selection by pathogens. By contrast, the simple elegance of the chicken MHC appears more the work of a single selective pressure which may or may not have anything to do with defence against pathogens⁵. Nevertheless, having crossed that particular road the chicken is seen to have put a lot of immunological eggs in a rather small basket. ■

Peter Parham is in the Departments of Structural Biology, and Microbiology and Immunology, Stanford University, Stanford, California 94305, USA.

1. Marsh, S. G. E., Parham, P. & Barber, D. L. *The HLA FactsBook* (Academic, London, 1999).
2. The MHC sequencing consortium *Nature* **401**, 921–923 (1999).
3. Parham, P. (ed.) *Immunol. Rev.* **167**, 1–379 (1999).
4. Kaufman, J. *et al.* *Nature* **401**, 923–925 (1999).
5. Hughes, A. L. & Hughes, M. K. *Nature* **377**, 391 (1995).
6. Skodt, K., Koch, C., Crone, M. & Simonsen, M. *Tissue Antigens* **25**, 278–282 (1999).
7. Briles, W., Briles, R., Taffs, R. & Stone, H. *Science* **219**, 977–979 (1983).
8. Brown, M. G., Scalzo, A. A., Matsumoto, K. & Yokoyama, W. *Immunol. Rev.* **155**, 53–65 (1997).
9. Wagtmann, N., Rojo, S., Eichler, E., Mohrenweiser, H. & Long, E. O. *Curr. Biol.* **7**, 615–618 (1997).
10. Uhrberg, M. *et al.* *Immunity* **7**, 739–751 (1997).
11. Biron, C. A., Byron, K. S. & Sullivan, J. L. *New Engl. J. Med.* **320**, 1731–1735 (1989).

Daedalus

Nuclear autumn

Global warming is generally blamed on the emission of greenhouse gases by human activities, chiefly the burning of hydrocarbons. But one theory blames the declining intensity of cosmic rays. As they traverse the atmosphere, they nucleate water vapour to cloud droplets, just as in a particle-detecting cloud chamber. Clouds reflect sunlight; so the lower the cosmic ray flux, the fewer clouds, the more sunlight hits the Earth, and the higher its temperature. So to counter global warming, we need more atmospheric radiation.

The nuclear industry will welcome this argument. Nuclear power not only reduces the need to burn hydrocarbons; it generates radioactive waste as well. If instead of storing the waste in careful seclusion, we released it as fine dust into the atmosphere, it would soon be wafted up to cloud altitudes. It would nucleate new clouds with splendid efficiency.

This simple proposal would arouse almost the ultimate in environmental outrage. But Daedalus reckons that conventional fuel-burning can do the job instead. It too releases particles into the atmosphere. The finest and most efficient nuclei are probably those released in annoying clouds by diesel engines, a growing segment of the automotive market. Many engineers are trying to modify the diesel engine to reduce its particulate emissions; but DREADCO engineers are modifying the fuel so as to optimize those emissions. Their aim is to reduce the size of the particles to much less than a wavelength of light. They will then be invisible, and will no longer annoy the public. They will also be ideal condensation nuclei.

The project has many hopeful clues to follow. Carbon black is made by burning fuel oil. For the finest soot particles, special alkali metal compounds are added to the oil. Fullerene chemists have their own tricks for burning fuels to very small carbon particles. So with good fortune DREADCO's clean-exhaust, high-nucleation, save-the-planet diesel fuel will soon hit the market. Ecologically aware consumers will rush to buy it. Its invisible exhaust particles, wafted aloft by atmospheric turbulence, will encourage global cloud cover and counter global warming. Sadly for cold, humid Britain, their nucleation properties will be even more effective at ground level. In dense winter traffic, the relentless nucleation of innumerable exhausts will create appalling fog on the motorways.

David Jones

Obituary

Arthur James Cain (1921–99)

The 1950s and 1960s saw a flowering of experiments on natural populations. Many evolutionary biologists tried to detect and measure natural selection, and to interpret evolution through the mathematics of R. A. Fisher, J. B. S. Haldane and Sewall Wright. The majority unequivocally supported a Darwinian view of the natural world. By the 1970s, however, uncertainty had set in. Studies of variation in proteins and DNA seemed to favour an alternative — that most evolutionary changes have little or no effect on survival or reproduction. A schism arose between ‘selectionists’ on one side and ‘neutralists’ on the other. The former often came from the ecological parts of the subject, and the latter from the molecular parts. The argument between them was long and hard, and still continues. Arthur Cain, who died on 20 August at the age of 78, was a standard-bearer for the selectionists.

Cain was educated at the Lawrence Sherrieff school, Rugby, England, from where he won an open scholarship to Magdalen College, University of Oxford. There, in the 1950s, with Philip Sheppard he carried out a classic study of land snails whose colour polymorphisms had been widely cited as neutral in terms of natural selection. In an elegant series of observations and experiments, the two collaborators showed that the different colours and patterns strongly affected the likelihood of being eaten by predators. This work convinced Cain that it is always a mistake to assume a character to be selectively neutral without a careful study of its interactions with the ecology and habits of the organism. Indeed, it is a double mistake because it discourages further observation and experiment.

These conclusions inspired him to accumulate evidence about the power and delicacy of natural selection. The result was a trenchant paper, ‘The Perfection of Animals’, first published in 1964 and reprinted in the *Biological Journal of the Linnean Society* in 1989. It is still necessary reading for nascent biologists. It is sad that Cain’s last illness prevented him from learning about the recent evidence for selection acting on supposedly ‘trivial’ variants of protein and DNA.

Following up the work with Sheppard, Cain joined John Currey to produce the first description of ‘area effects’, parts of the range of a species where particular genetic variants are found at high frequencies in many adjacent populations,



Naturalist and doughty Darwinian

apparently regardless of habitat, separated from other such groups of populations by steep gradients in gene frequency. In 1963, Cain and Currey argued that these patches were caused by cryptic environmental selection. The work has given rise to much argument in the realms of phylogeography.

Cain’s interest in evolution was wide. An early expedition to the Solomon Islands, and later ones to British Guiana (as it then was) and the Dominican Republic, reinforced his fascination with the processes by which species come about. He produced a seminal book, *Animal Species and Their Evolution* (1954, reprinted by Princeton University Press in 1993), and worked on the evolution of fruit pigeons, parrots, ducks, earthworms, fruit flies, sea anemones, microbes and plants. He developed an extraordinary breadth and depth of knowledge about animals and plants, and an uncanny acuteness of biological understanding. He was, in the very best sense of that word, a naturalist, perhaps the most accomplished of his day.

Early in the 1960s, Cain also became a leader in the theory of classification, and a pioneer of numerical methods in taxonomy, although his approach became superseded by that devised by Robert Sokal and Peter Sneath. His work on speciation led him to investigate the history of the species concept, and he grew interested in Linnaeus, about whom he wrote several papers. He branched out into other aspects of scientific history, ranging from Aristotle to Darwin.

Cain loved to campaign as much as he loved to argue. He had an incendiary enthusiasm for evolutionary biology, a tendency to wrathful indignation, and a deep contempt for administrators, particularly those who had once been scientists. When in 1985 he was awarded the Linnean Society’s Gold Medal, he took the opportunity to attack the current vogue for appraisal and assessment. When a paper from his doctoral thesis on the staining of lipids became a ‘citation classic’, he responded by writing a thinly veiled attack on quantitative measures of merit.

After Oxford and four years in Manchester, Cain moved to the University of Liverpool in 1968. When he became chairman of the Department of Zoology there, he circulated an edict that members of staff should come to his office or laboratory only to discuss science. Matters of administration would be settled over morning coffee or afternoon tea. He wrote a book on Erasmus Darwin that never saw the light of day, for reasons very much in character. Apparently he submitted the manuscript to a well-known publisher, whose editor said that they would be delighted to accept it, but that it contained too many quotations. Would Cain care to reduce their number? He would not. In high dudgeon he submitted it to another publisher, with the same result. After that he became exasperated and decided not to publish it at all. He put it in the university library at Liverpool, where it still lies.

His combative style and changeable moods made Cain some enemies as well as friends — which perhaps accounts for his rather late election to the Royal Society in 1989. On his good days he could charm anybody and on his bad days he could offend anybody. Nonetheless he gained the lasting respect, affection and gratitude of many students and colleagues. He had a sort of passionate innocence that was very appealing. At a meeting in the 1980s it was estimated that about 120 (40%) of those attending were his academic descendants (students, or students of students, or students of students of students). Their company included many figures who are now prominent in evolutionary biology. Few of them will forget Arthur Cain’s unique combination of enthusiasm, encyclopaedic knowledge, intelligence, encouragement, irascibility and disapproval. That particular mixture had a lasting influence on the growth of his subject.

Bryan Clarke

Bryan Clarke is in the Institute of Genetics, Queen’s Medical Centre, Clifton Boulevard, Nottingham NG7 2UH, UK.
e-mail: Bryan.Clarke@nottingham.ac.uk

The colourful world of the mantis shrimp

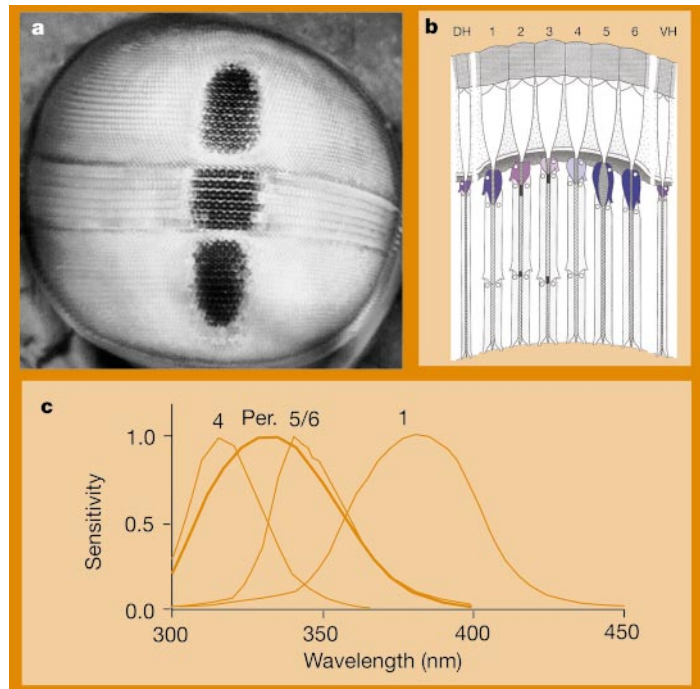
The colour-vision system of these crustaceans includes four types of UV photoreceptor.

Humans cannot see ultraviolet light, but many arthropods and vertebrates can because they have a single photoreceptor with a peak sensitivity to light at wavelengths of around 350 nanometres (ref. 1). Here we use electrophysiological methods to investigate the vision of the mantis shrimp, *Neogonodactylus oerstedii*. We find that this marine crustacean has at least four types of photoreceptor for ultraviolet light that are located in cells of the eye known as R8 cells. These photoreceptors are maximally sensitive to light of wavelengths 315, 330, 340 and 380 nm. Together with previous evidence², this finding indicates that the remarkable colour-vision system in these stomatopod crustaceans may be unique, as befits their habitat of kaleidoscopically colourful tropical coral reefs.

Many stomatopods inhabit the top few metres of water, which are bathed in light of ultraviolet-A wavelengths³. In this spectrally varied environment, they have evolved extremely complex retinæ within their apposition compound eyes^{2,4-6}. The top four rows of the midband (Fig. 1a) contain eight spectral sensitivities from wavelengths of 400 to 700 nm, each based on a different visual pigment². Oversampling within the spectrum is avoided by elegant filtering mechanisms that narrow these sensitivities^{2,6}.

The R8 cells of stomatopods, like the cells R1 to R7 (see Fig. 1), have an extreme proliferation of spectral sensitivities. Three of the four ultraviolet spectral sensitivities are found in the midband, and one is in the

Figure 1 R8 photoreceptors in the midband of the stomatopod eye have multiple UV sensitivities **a**, Stomatopod eye showing the clearly defined midband, rows 1 to 6 (dorsal to ventral)^{4,5}. The top four rows are concerned with colour information, and the remaining two are specialized for polarization^{2,5}. **b**, Diagram of the photoreceptors in the six midband rows and the peripheral retinae. R8 cells are coloured. Chromatic channels from 400 to 700 nm are contained in a population of cells called R1 to R7 (ref. 4). VH and DH are representatives of the dorsal and ventral hemisphere regions in the peripheral retina. **c**, Ultraviolet spectral sensitivities of R8 cells. Per, peripheral retina (DH or VH).



periphery. Two of these sensitivities are relatively broad and two are very narrow, but they are all too narrow to be the result of visual-pigment absorption alone⁷. Filtering mechanisms like those in the rest of the eye are almost certainly the reason for this spectral tuning^{2,6}, with the filter being situated in the dioptric apparatus or within the microvilli of the R8 photoreceptors^{8,9}. Midband row 2 or row 3 R8 spectral sensitivities

have not yet been characterized, as they are hard to find or record from.

The evenly spaced ultraviolet photoreceptors in the eye of *N. oerstedii* indicate that stomatopods have extended the high-frequency sampling system of the range 400 to 700 nm into the ultraviolet to form a single colour-vision system that is sensitive from 300 to 700 nm.

The narrowed sensitivities of all four cells can be modelled using existing visual-pigment nomograms⁷ and a series of corneal ultraviolet cut-off filters¹⁰ (Fig. 2). It is not possible to generate the spectral sensitivities observed by differential filtering of the same visual pigment. Therefore, in common with the longer-wavelength sensitivities in the R1 to R7 portion of the eye, ultraviolet spectral sensitivities of the R8 cells seem to be generated from a population of different visual pigments¹⁰.

For colour vision, narrow-band photoreceptors have several advantages over broad-band systems, including fine colour discrimination¹¹ and improved colour constancy^{11,12}. The main drawback is a drastic reduction in sensitivity⁶, but many stomatopods are found in the brightly lit waters of coral reefs and are usually diurnally active, so highly sensitive vision is not required⁷.

Stomatopods use colour in communication¹³ and possess colour vision¹⁴ but it has remained unknown whether these extend into the ultraviolet as they do in birds and

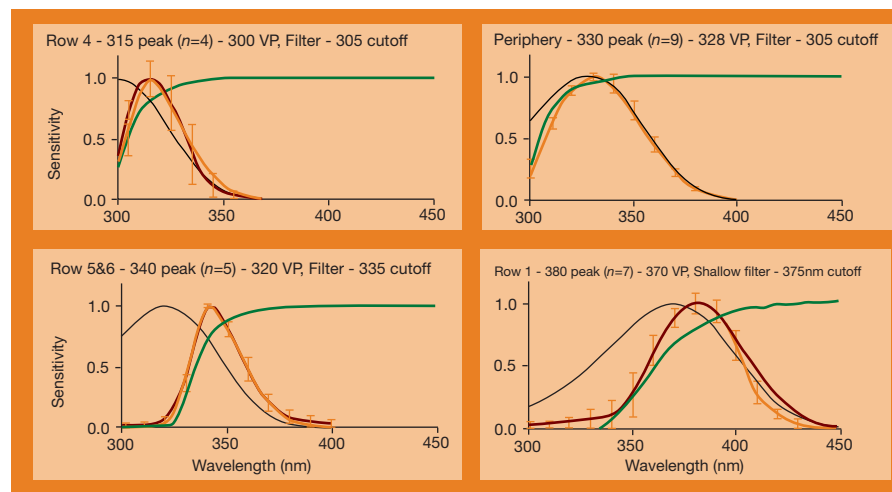


Figure 2 The multiple ultraviolet sensitivities are caused by several different rhodopsins that are heavily filtered. Electrophysiological spectral sensitivities (orange) are shown with standard errors for number of cells measured (n). A modelled spectral sensitivity (thick black line) has been fitted to these. Model components are possible visual-pigment absorbances calculated from ref. 7 (thin black lines) and putative filters that narrow the visual-pigment absorbance (green lines). Filter spectra are chosen from a series of known ultraviolet filters (J.M. and U. Siebeck, unpublished data) with 50% cut-off points in the range 300 to 400 nm. These intracellular recordings from stomatopod photoreceptors and our recordings from R1–7 cells in the range 400 to 700 nm (data not shown) broadly confirm previous microspectrophotometric results^{2,6}.

fish¹, although many coloured areas used by stomatopods in communication contain ultraviolet components¹⁴. Most of the colour information from natural objects can be decoded by just four types of photo-receptor in the 300–700 nm range^{12,15}, so there must be other reasons for the bizarre retinal design of stomatopods, which probably uses 12 channels for colour. There are two possible explanations that logically extend the sensitivity range of 400 to 700 nm.

The first possibility is that stomatopod eyes examine colour space from 300 to 700 nm in much the same way as the ear examines auditory space. The multiple, narrow-band spectral sampling channels, from 300 to 700 nm, may be analogous to the different auditory frequencies to which a cochlea is tuned along its length. This may be thought of as a kind of 'digital' colour vision.

Alternatively, stomatopods may divide the spectral world from 300 to 700 nm into six dichromatically examined windows (two in the ultraviolet, mediated by row 1–4 R8 cells), each of which is subject to very fine spectral discrimination¹⁴.

One of the ultraviolet sensitivities is in row 5 of the midband, a region of the eye that is probably used in polarization vision^{4,5}. Anatomical⁴ and microspectrophotometry⁶ comparisons show that the same sensitivity

exists in row 6 and that this is theoretically optimal for polarized-light vision¹⁶.

As ultraviolet-blind humans, we have created a barrier in the spectrum at 400 nm. For the colour or polarization vision of stomatopods, this barrier is meaningless.

Justin Marshall*, Johannes Oberwinkler†

*VTHRC, University of Queensland, Brisbane 4071, Australia

†Department of Neurobiophysics, University of Groningen, Nijenborgh 4, 9747 AG, Groningen, The Netherlands

1. Tovee, M. J. *Trends Ecol. Evol.* **10**, 455–460 (1995).
2. Cronin, T. W. & Marshall, N. J. *Nature* **339**, 137–140 (1989).
3. Smith, R. C. & Baker, K. S. *Appl. Optics* **20**, 177–184 (1981).
4. Marshall, N. J. *Nature* **333**, 557–560 (1988).
5. Marshall, N. J. *et al. Phil. Trans. R. Soc. Lond. B* **334**, 33–56 (1991).
6. Cronin, T. W. *et al. Vision Res.* **34**, 2639–2656 (1994).
7. Hart, N. S., Partridge, J. C. & Cuthill, I. C. *J. Exp. Biol.* **201**, 1433–1446 (1998).
8. Hardie, R. C. *Trends Neurosci.* **9**, 419–423 (1986).
9. Arikawa, K. *et al. Vision Res.* **39**, 1–8 (1999).
10. Cronin, T. W. & Marshall, N. J. *J. Comp. Physiol. A* **166**, 261–275 (1989).
11. Neumeier, C. in *Vision and Visual Dysfunction: Evolution of the Eye and Visual System* Vol. 2 (eds Cronin-Dillon, J. R. & Gregory, R. L.) 284–305 (Macmillan, London, 1991).
12. Osorio, D., Marshall, N. J. & Cronin, T. W. *Vision Res.* **37**, 3299–3309 (1997).
13. Caldwell, R. L. & Dingle, H. *Sci. Am.* **234**, 80–89 (1976).
14. Marshall, N. J., Jones, J. P. & Cronin, T. W. *J. Comp. Physiol. A* **179**, 473–481 (1996).
15. Vorobyev, M., Osorio, D., Bennett, A. T. D. & Cuthill, I. C. *J. Comp. Physiol. A* **183**, 621–633 (1998).
16. Seliger, H. H., Lall, A. B. & Biggley, W. H. *J. Comp. Physiol. A* **175**, 475–486 (1994).

AIDS

Re-emergence of HIV after stopping therapy

A dormant reservoir of human immunodeficiency virus (HIV) is established early on during primary infection¹ which consists of latently infected, resting CD4⁺ T cells carrying replication-competent HIV. This pool can persist even in individuals who are receiving highly active antiretroviral therapy (HAART)^{2–4}. Here we show that this pool rapidly re-emerges within weeks of discontinuing HAART in two patients, and that this re-emergence is associated with the appearance of HIV in the plasma (viraemia) of these patients. Both had been aviraemic while receiving HAART and

intermittent treatment with interleukin-2 (ref. 5), and repeated attempts to isolate replication-competent HIV in this population of cells during therapy had been unsuccessful. This finding raises the possibility that there may be other tissue reservoirs of HIV that contribute to early plasma viral rebound following discontinuation of HAART in infected patients.

Despite the success of HAART in driving plasma viraemia to below the levels of detectability in many HIV-infected individuals^{6,7}, the persistence of a latent reservoir of HIV in resting CD4⁺ T cells in HAART-treated individuals is considered to be a major impediment to the long-term control of HIV infection⁸. We have recently shown that intermittently administering interleukin-2 during continuous HAART reduces the size of this reservoir in resting CD4⁺ T cells⁵. The only way to demonstrate that cellular reservoirs of HIV have been eradicated is by clinical trial in which therapy is discontinued in individuals who have become aviraemic with therapy, so we enrolled two patients from a previously reported cohort⁵ in such a trial. This trial, involving 18 patients, will be the subject of a separate report⁹.

Before HAART was discontinued in these two patients, we were unable to detect replication-competent HIV in the resting CD4⁺ T cells of peripheral blood at two

consecutive time points, despite culturing up to 330 million cells⁵, and the virus could not be isolated from the lymph nodes of either patient⁵. After HAART was discontinued, we were able to detect plasma viraemia within three weeks in both patients.

We carried out quantitative co-culture assays^{5,10} by using highly purified resting CD4⁺ T cells at several time points. The pool of resting CD4⁺ T cells carrying replication-competent HIV emerged shortly after plasma viraemia was detected (Fig. 1). This pool of cells increased in size by 3.8 log by week 4 and by 2.0 log by week 6 in patients 1 and 13 (numbered according to ref. 5), respectively, following the discontinuation of HAART. The integrated form of HIV DNA, which was not detected by Alu-LTR polymerase chain reaction (Table 1) in 10⁶ resting CD4⁺ T cells from either patient during HAART, was readily detectable during the reappearance of the virus in both patients (142 copies per million cells for patient 1 at week 6, and 420 copies for patient 13 at week 8).

Our results demonstrate the speed with which the latent HIV reservoir in the resting CD4⁺ T-cell compartment emerges during the reappearance of plasma viraemia following discontinuation of HAART. This is as fast as the initial establishment of the reservoir in patients during primary infection¹. Furthermore, given that replication-competent HIV was not detectable in resting CD4⁺

Table 1 Quantitative analysis of integrated HIV-1 DNA in resting CD4⁺ T cells

Patient	Integrated HIV-1 proviral load (copies per 10 ⁶ cells)	
	Week 0	Week 6
1	<0.5	142
	Week 0	Week 8
13	<0.5	420

Genomic DNA from resting CD4⁺ T cells was serially diluted and subjected to the polymerase chain reaction (PCR) in duplicate using nested 5' primer from conserved Alu and 3' primer from conserved HIV-1 long-terminal repeat (LTR) sequences as described⁵. A portion of the diluted first PCR product was further subjected to the second round of PCR using nested HIV-1 LTR-specific primers. The product of two rounds of PCR was detected by dot blotting using LTR-specific probe. Week 0 represents the time when patients discontinued HAART.

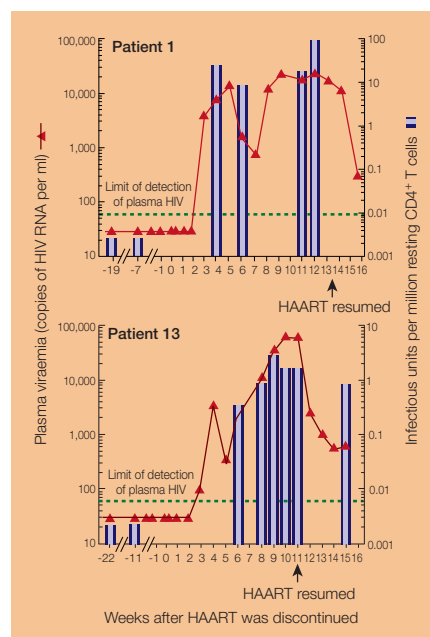


Figure 1 Rebound in plasma viraemia and re-emergence of the pool of latently infected, resting CD4⁺ T cells in two patients after HAART was discontinued. Plasma HIV RNA was measured by using the bDNA assay (Chiron) with a detection limit of 50 copies per ml. Frequencies of resting CD4⁺ T cells carrying replication-competent HIV were determined by quantitative co-culture assays as described^{5,10}. Patients 1 and 13 received HAART for 33 and 30 months, respectively, before it was discontinued. Week 0 represents the time when patients discontinued HAART.

T cells from either the lymph nodes or peripheral blood before HAART was discontinued in these patients, it is possible that other tissue reservoirs are responsible for the rebound in plasma viraemia. Genotypic and phenotypic studies will be required to determine the source of this rebounding virus.

Tae-Wook Chun, Richard T. Davey Jr, Delphine Engel, H. Clifford Lane, Anthony S. Fauci

Laboratory of Immunoregulation,
National Institute of Allergy and Infectious
Diseases, National Institutes of Health,
Bethesda, Maryland 20892, USA
e-mail: twchun@nih.gov

1. Chun, T. W. *et al.* *Proc. Natl Acad. Sci. USA* **95**, 8869–8873 (1998).
2. Chun, T. W. *et al.* *Proc. Natl Acad. Sci. USA* **94**, 13193–13197 (1997).
3. Finzi, D. *et al.* *Nature Med.* **5**, 512–517 (1999).
4. Zhang, L. *et al.* *N. Engl. J. Med.* **340**, 1605–1613 (1999).
5. Chun, T. W. *et al.* *Nature Med.* **5**, 615–655 (1999).
6. Gulick, R. M. *et al.* *N. Engl. J. Med.* **337**, 734–739 (1997).
7. Hammer, S. M. *et al.* *N. Engl. J. Med.* **337**, 725–733 (1997).
8. Cohen, O. J. & Fauci, A. S. *J. Am. Med. Assoc.* **280**, 87–88 (1998).
9. Davey, R. T. *et al.* *Proc. Natl Acad. Sci. USA* (in the press).
10. Chun, T. W. *et al.* *Nature* **387**, 183–188 (1997).

Statistical mechanics

Microscopic chaos from brownian motion?

Gaspard *et al.*¹ have analysed a time series of the positions of a brownian particle in a liquid, and claimed that it provides empirical evidence for microscopic chaos on a molecular scale. An accompanying comment² emphasized the fundamental nature of the experiment. Here we show that virtually identical results can be obtained by analysing a corresponding numerical time series of a particle in a manifestly microscopically non-chaotic system.

Like Gaspard *et al.*¹, we have analysed the position of a single particle colliding with many others. We used the Ehrenfest wind-tree model³, in which the point-like ('wind') particle moves in a plane, colliding with randomly placed, fixed square scatterers ('trees') (Fig. 1a). We chose this model because collisions with the flat sides of the squares do not lead to exponential separation of corresponding points on initially nearby trajectories, so there are no positive Lyapunov exponents, which are characteristic of microscopic chaos. In contrast, Gaspard *et al.*¹ used a Lorentz model as being similar to brownian motion, a model in which the squares are replaced by hard, circular discs. This does exhibit exponential separation of nearby trajectories, leading to a positive Lyapunov exponent and hence microscopic chaos.

Nevertheless, despite being non-chaotic, the Ehrenfest model reproduces all the results presented by Gaspard *et al.*¹. The particle trajectory segment shown in Fig. 1b

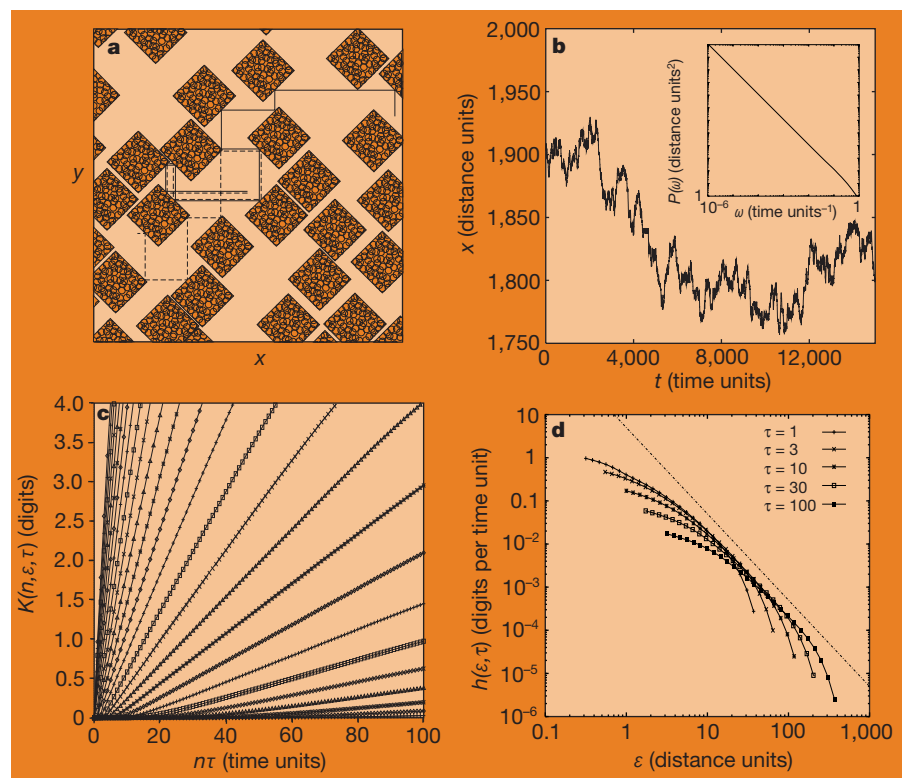


Figure 1 Brownian motion results of Gaspard *et al.*¹ numerically reproduced from the non-chaotic Ehrenfest wind-tree model (notation as in ref. 1). The square scatterers have a diagonal of two length units and fill half the area considered. The particle moves with unit velocity in four possible directions. The position on its trajectory is determined for 10^6 points separated by one time unit. **a**, Two nearby trajectories split only at a corner C; no exponential separation occurs (see Fig. 1 of ref. 1). **b**, A typical trajectory is diffusive with an ω^{-2} power spectrum (inset), where ω is the angular frequency (see Fig. 2 of ref. 1). **c**, The information entropy $K(n, \epsilon, \tau)$ for $\tau=1$ and $\epsilon=0.316 \times 1.21^m$, where m is an integer running from 0 to 25 (see Fig. 3 of ref. 1). **d**, The envelope of the slopes of these K curves, $h(\epsilon, \tau)$ appears to imply a positive (chaotic) h_{KS} for the Ehrenfest model, as for brownian motion (see Fig. 4 of ref. 1).

is strikingly similar to that for the brownian particle (Fig. 2 of Gaspard *et al.*). Our subsequent analysis parallels that of Gaspard *et al.*¹, where further details may be found.

The microscopic 'chaoticity' is determined by estimating the Kolmogorov–Sinai entropy h_{KS} as described^{4,5} by using the information entropy $K(n, \epsilon, \tau)$ obtained from the frequency with which the particle retraces part of its (previous) trajectory within a distance ϵ , for n measurements spaced at a time interval τ . As h_{KS} here equals the sum of the positive Lyapunov exponents, the determination of a positive h_{KS} would imply microscopic chaos. Like Gaspard *et al.*¹, we find that K grows linearly with time (Fig. 1c), giving a positive (non-zero) bound on h_{KS} (Fig. 1d). Indeed, Fig. 1b–d for a microscopically non-chaotic model are virtually identical to the corresponding Figs 2–4 of Gaspard *et al.* Thus, Gaspard *et al.* did not prove the presence of microscopic chaos in brownian motion.

The algorithm of refs 4,5 as applied here cannot determine the microscopic chaoticity of brownian motion because the time interval between measurements, 1/60 s (ref. 1), is so much larger than the microscopic timescale determined by the inverse collision frequency in a liquid, which is approximately 10^{-12} s. A decisive determination of

microscopic chaos would require a time interval τ of the same order as characteristic microscopic timescales.

C. P. Dettmann*, E. G. D. Cohen*, H. van Beijeren†

*Center for Studies in Physics and Biology,
Rockefeller University, New York,
New York 10021, USA

†Institute for Theoretical Physics,
University of Utrecht,
3584 CC Utrecht, The Netherlands

1. Gaspard, P. *et al.* *Nature* **394**, 865–868 (1998).
2. Durr, D. & Spohn, H. *Nature* **394**, 831–833 (1998).
3. Ehrenfest, P. & Ehrenfest, T. *The Conceptual Foundations of the Statistical Approach in Mechanics* (transl. Moravcsik, M. J.) 10–13 (Cornell Univ. Press, Ithaca, New York, 1959).
4. Grassberger, P. & Procaccia, I. *Phys. Rev. A* **28**, 2591–2593 (1983).
5. Cohen, A. & Procaccia, I. *Phys. Rev. A* **31**, 1872–1882 (1985).

Gaspard *et al.*¹ have shown that the position of a brownian particle behaves like a Wiener process with positive resolution-dependent entropy². More surprisingly^{3–5}, they claim that this observation provides proof of 'microscopic chaos', a term they illustrate by examples of finite dimensional dynamical systems which are intrinsically unstable. We do not believe that they have provided evidence for microscopic chaos in the sense in which they use the term.

Although the recent literature finds such

chaos on a molecular level quite plausible, the observed macroscopic disorder cannot be taken as direct evidence of microscopic chaos. The effectively infinite number of molecules in a fluid can generate the same macroscopic disorder without any intrinsic instability, so brownian motion can be derived for systems that would usually be called non-chaotic, such as a tracer particle in a non-interacting ideal gas. All that is needed for diffusion is 'molecular chaos' in the sense of Boltzmann, that is, the absence of observable correlations in the motion of single molecules.

Part of the confusion is due to the lack of a unique definition of microscopic chaos for systems with an infinite number of degrees of freedom. Gaspard *et al.*¹ introduced the term by extrapolating from finite dimensional dynamical systems for which chaos is well defined: on average, initially close states separate exponentially when time tends to infinity. The Lyapunov exponent, or rate of separation, is independent of the particular method used to measure 'closeness'. However, the ideas of diffusion and brownian motion involve infinitely many degrees of freedom. In this thermodynamic limit, Lyapunov exponents are no longer independent of the metric. The large system limit of a finite non-chaotic system will therefore remain non-chaotic with one particular metric and become chaotic with another.

We can illustrate this point by using an example⁶ introduced in the context of cellular automata. Consider two states $x = (\dots x_{-2}, x_{-1}, x_0, x_1, x_2, \dots)$ and $y = (\dots y_{-2}, y_{-1}, y_0, y_1, y_2, \dots)$ of a one-dimensional bi-infinite lattice system. If the distance between x and y is defined by $d_{\max}(x, y) = \max_i |x_i - y_i|$, it can grow exponentially only if the local differences do. This is what is usually meant by 'chaos', and what Gaspard *et al.*¹ meant by 'microscopic chaos'. This mechanism is absent in the thermodynamic limit of finite non-chaotic systems, so the limit could also be said to be non-chaotic⁷⁻⁹. However, the distance $d_{\exp}(x, y) = \sum_i |x_i - y_i| e^{-|i|}$ can also show exponential divergence if an initially distant perturbation moves towards the origin without growing⁶. When observing a localized tracer, as Gaspard *et al.*¹ did, the latter choice may be the more appropriate.

In finite dimensional dynamical systems, chaos arises as a result of the defocusing microscopic dynamics. Positive entropy is generated by the initially insignificant digits of the initial condition brought forth by the dynamics. At the thermodynamic limit, a different mechanism also exists: perturbations coming from distant regions kick the tracer particle once and move away again to infinity. The entropy is positive because of information stored in remote parts of the initial condition. Suitable Lyapunov exponents⁶ can be defined for this case as well.

To resolve the confusion, we propose letting the system size tend to infinity first, before the observation time¹⁰. A system observed in a particular metric μ is then described as μ -chaotic when we find a positive Lyapunov exponent using this metric. Gaspard *et al.*¹ had in mind the type of chaos that is detectable with the metric $d_{\max}$ and arises from a local instability. But they fall short of providing experimental evidence for this specific type of chaos.

Peter Grassberger, Thomas Schreiber

Department of Physics, University of Wuppertal,
42097 Wuppertal, Germany

1. Gaspard, P. *et al.* *Nature* **394**, 865–868 (1998).
2. Gaspard, P. & Wang, X.-J. *Phys. Rep.* **235**, 291–345 (1993).
3. Dürr, D. & Spohn, H. *Nature* **394**, 831–833 (1998).
4. Schewe, F. & Stein, B. *Physics News Update* <<http://www.aip.org/enews/physnews/1998/split/pnu389-3.htm>> (1998).
5. Rauner, M. *Physik. Blätt.* **54**, 1001 (1998).
6. Wolfram, S. *Physica D* **10**, 1–35 (1984).
7. Crutchfield, J. & Kaneko, K. *Phys. Rev. Lett.* **60**, 2715–2718 (1988).
8. Ershov, S. V. & Potapov, A. B. *Phys. Lett. A* **167**, 60–64 (1992).
9. Politi, A., Livi, R., Oppo, G.-L. & Kapral, R. *Europhys. Lett.* **22**, 571–576 (1993).
10. Sinai, Y. G. *Int. J. Bifurcat. Chaos* **6**, 1137–1142 (1996).

Gaspard et al. reply — We presented experimental evidence for the existence of microscopic chaos in the mesoscopic motion of a brownian particle in solution. We used standard techniques to analyse long trajectories of a brownian particle and inferred a dynamical entropy from this analysis. We showed that this dynamical entropy can be accessed experimentally by measuring multiple-time correlation functions for the particle. The dynamical entropy was then used to provide a positive lower bound for the sum of the Lyapunov exponents for the underlying deterministic dynamical system composed of the fluid and brownian particles. We concluded that the positive dynamical entropy, obtained experimentally, was evidence for the existence of positive Lyapunov exponents in the underlying dynamics, and hence for the existence of microscopic chaos generated by a dynamical instability.

Dettmann *et al.* and Grassberger and Schreiber argue that there may be other explanations. There is no doubt that a positive dynamical entropy could be generated by mechanisms other than a local intrinsic, dynamical instability. Illustrative models include the Rayleigh flight of a tracer particle in a non-interacting ideal gas, the motion of an impurity in a harmonic crystal, and the motion of the wind particle in the wind-tree model. In these models, an external pseudorandom generator must be involved at some stage of the simulation of the dynamics, leading to a positive dynamical entropy without a local dynamical instability.

In the Rayleigh flight and the harmonic crystal, randomness is generated repetitively by the new particles or waves continuously

coming from infinity. In the wind-tree model, the scatterers are randomly located by using a pseudorandom generator before the simulation of the time evolution. As the wind particle collides with a new scatterer, it picks up new information on the location of this scatterer and the information recorded on this trajectory grows as a result. However, if the square scatterers formed a regular lattice, this growth would stop and the dynamical entropy would vanish. In contrast, for circular disks, as in the Lorentz gas, the dynamical entropy is always positive whether the lattice is regular or not. The apparent dynamical randomness of the wind-tree model therefore has its origin in the structural disorder of the model.

The models mentioned by Dettmann *et al.* and Grassberger and Schreiber fail to capture an essential feature of brownian motion: that the diffusion coefficient is inversely proportional to the viscosity of the fluid surrounding the brownian particle. The viscosity of the fluid is absent in models in which the surrounding particles either do not move or have no interaction or only harmonic interactions. These models are therefore not plausible for the interpretation of our experiment¹.

However, if the interactions between surrounding particles are generic and nonlinear, the viscosity can be positive and the dependence of the diffusion coefficient on the viscosity can be explained. In such systems with nonlinearly interacting particles, numerical and analytical studies have shown that intrinsic local instability is the dominant mechanism for generating dynamical randomness, and no external pseudorandom generator is needed^{2,3}. Standard models of brownian motion are of this kind and develop chaotic dynamics with a full spectrum of positive Lyapunov exponents.

For these reasons, we believe that our explanation in terms of a randomness self-generated by the dynamical instability is the most plausible one for this experiment. We hope that further experiments will be able to distinguish more directly between different regimes of random behaviour in systems with a large number of particles, and so provide further evidence of microscopic chaos.

P. Gaspard*, M. E. Briggs†, M. K. Francis‡, J. V. Sengers‡, R. W. Gammon‡, J. R. Dorfman‡, R. V. Calabrese‡

*Faculté des Sciences, Université Libre de Bruxelles,
Boulevard du Triomphe, 1050 Brussels, Belgium

†University of Utah,
Salt Lake City, Utah 84112, USA

‡University of Maryland,
College Park, Maryland 20742, USA

1. Gaspard, P. *et al.* *Nature* **394**, 865–868 (1998).
2. Posch, H. A. & Hoover, W. G. *Phys. Rev. A* **38**, 473–482 (1988).
3. van Beijeren, H., Dorfman, J. R., Posch, H. A. & Dellago, C. *Phys. Rev. E* **56**, 5272–5277 (1997).

Inferring the historical patterns of biological evolution

Mark Pagel

School of Animal and Microbial Sciences, University of Reading, Whiteknights, Reading RG6 6AJ, UK

Phylogenetic trees describe the pattern of descent amongst a group of species. With the rapid accumulation of DNA sequence data, more and more phylogenies are being constructed based upon sequence comparisons. The combination of these phylogenies with powerful new statistical approaches for the analysis of biological evolution is challenging widely held beliefs about the history and evolution of life on Earth.

Since 1981, the number of articles reporting phylogenies based on gene-sequence information has been increasing exponentially, with more than half appearing since 1996 (Fig. 1), as new sequence information becomes available at exhausting rates. The phylogenies span taxonomic groups ranging from viruses to bacteria, fungi, plants and animals. The prospect of describing in detail the patterns of descent within many of the major groups of organisms, seen as fanciful just ten years ago, is now realistic.

However, the influence of these new phylogenies extends beyond cataloguing the relatedness of species. All the events of biological evolution are played out somewhere along the branches of phylogenetic trees. As a consequence, these phylogenetic trees preserve traces of the historical evolutionary processes that gave rise to the diversity of contemporary species. This raises the intriguing possibility that the combination of a phylogeny and information on species can be used to infer what the past was like and how the present came about.

I shall not be concerned here with the reconstruction of phylogenies themselves, as this has received much attention (see, for example, ref. 1). Rather, I shall describe the recent advances in statistical modelling of evolution on phylogenetic trees, particularly the use of maximum-likelihood techniques, that are providing researchers with new ways to investigate the evolution of life on Earth. What sets the new statistical techniques apart from conventional non-statistical methods or palaeontological approaches is the range of characteristics that can be investigated and the nuances of historical trends of evolution that can be characterized. The result is that a new era of biological studies is emerging in which statistical approaches applied to phylogenies and information about species form an independent branch of historical enquiry.

Four areas of historical evolutionary enquiry have benefited most from the new statistical approaches: reconstruction of ancestral

character states, using a phylogeny in combination with a statistical description of how the traits of organisms evolve, to discover the most probable characteristics of ancestral species; estimation of the timings of historical evolutionary events; assessment of the tempo of evolution, which in turn allows the testing of punctuational or gradual models of evolution; and comparative studies, using correlation and regression to investigate which features of organisms change with which other features or with aspects of their environment, which can provide evidence for the temporal order of changes in two traits, suggesting probable causal pathways.

To illustrate the progress being made in these four areas, I draw on recent studies in which statistical models applied to phylogenies of organisms have been used to infer unexpected features of the common ancestor to life, to challenge conventional views about the Cambrian explosion and the effects of the Cretaceous–Tertiary (K–T) extinction, to reconstruct ancient proteins, to investigate gene–culture coevolution in human societies, and to test widely held theories of mammalian brain-size evolution.

Statistical models

The motivation for using statistical models to infer historical patterns of evolution lies in the belief that the diversity of contemporary species reflects the action of various evolutionary processes², including the rate and tempo of evolution, timings, correlations and ancestral states. A statistical model specifies by its parameters the way in which a process unfolds over time, or how a past feature leaves traces in the present. This is equivalent to identifying the contemporary signature or imprint of historical events.

The statistical parameters of a model of evolution are typically estimated by maximum-likelihood methods, in which the observed data on species and the model are represented in a common probabilistic framework. Likelihood methods regard the observed data as a fixed observation and seek the values of the statistical parameters that provide the most probable description of those data, given the model of evolution³. The likelihood does not describe either the probability that the events under study happened (they did) or that the model is true. Rather, it describes the likelihood that a given process as opposed to some other is responsible for the observed data. These properties make likelihood particularly suited to historical inference problems, in which the observed data arise only once.

If the model of evolution is a hypothesis to explain the data, likelihood chooses the hypothesis that best fits those data³. Hypotheses (models) are tested by the likelihood ratio statistic (LR), defined as $LR = -2\log_e[H_1/H_2]$, where, by convention, H_1 is the likelihood associated with the hypothesis that fits the data less well. If the hypotheses are special cases of one another, then LR approximates to a χ^2 statistic with degrees of freedom equal to the difference in the number of free parameters in the two models.

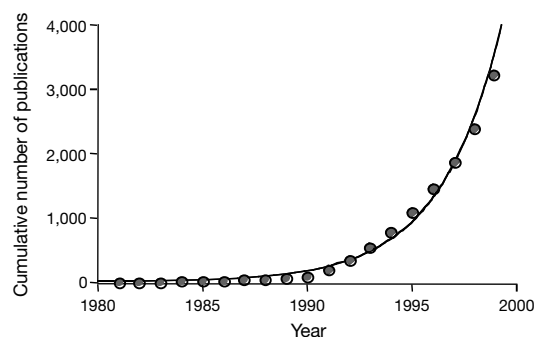


Figure 1 Cumulative number of publications in the Science Citation Index since 1981 that cite the terms 'molecular' and 'phylogeny' in the key words or abstract. The actual number of articles reporting gene-sequence-based phylogenies will be higher. The value for 1999 is extrapolated from the first seven months' data. The fitted curve accounts for more than 97% of the variance and has a period doubling time of just over two years.

Otherwise, LR greater than 4.0 is conventionally taken as evidence that one of the two hypotheses explains the data significantly better than the other³.

I shall assume here that the phylogenetic tree is known independently and without error, and shall focus instead on how phylogenies can be used to reveal the past. Nevertheless, it may be useful to define some terminology of phylogenetic trees: the root is the common ancestor to all the species in a phylogenetic tree; branches emanate from the root, tracing the course of evolution to descendants; these descendants reside at nodes of the tree; branches emanate from these nodes, eventually reaching the tips, or the contemporary organisms. Branches have lengths, typically measured in units of time or genetic divergence. A lineage can be loosely defined as a given pathway from the root of the phylogeny along connected branches to some ancestral node or contemporary species. See Box 1 for an explanation of some statistical terms.

Reconstructing ancestral states

Reconstructions of ancestral character states make it possible in principle to describe what the past was like, to discover how traits evolve and to understand function. They are increasingly being used to reconstruct proteins and genes that existed millions of years ago^{4,5}, and may provide insight into how genes will respond when subjected to forced-evolution methods. New statistical models incorporate explicit models of trait evolution with, amongst other things, the capability to detect directional trends, thereby raising the possibility of reconstructing ancestral features outside the range of features observed in the data.

The Markov-transition process is the statistical model widely used to describe the evolution of traits that adopt only a finite number of states. It is routinely used in phylogeny reconstruction¹ and in comparative methods^{2,6}. More recently, it has been applied to reconstructing ancestral character states on phylogenies^{7–12}, including protein evolution, studies of sexual selection, and habit and diet preferences⁸.

The Markov approach estimates the rates at which a discrete character makes transitions among its possible states as it evolves through time. These rates are sufficient to calculate the most probable states at ancestral nodes of the phylogeny. The maximum-likelihood estimate of an ancestral state is determined by independently calculating the likelihood of observing the species data, having successively fixed the particular ancestral node at the possible states of the character; the state with the largest likelihood corresponds to the most probable state at that node¹².

Maximum parsimony is the principal alternative method to likelihood reconstructions of ancestral states of discrete characters, and is still the more widely used. Parsimony reconstructions choose the state of the ancestor to minimize the number of evolutionary events required to explain the character states of the species. Parsimony works well when change is rare or branches are short (see, for example, ref. 13), because, under these circumstances, the probability of a character changing in a branch of the phylogenetic tree is not strongly related to its length. But parsimony methods can perform poorly, especially when rates of character evolution are high and the phylogeny includes some long branches^{1,14,15}. This is because parsimony does not take into account the lengths of the branches, so it has a tendency to underestimate the amount of change in long branches. Under some other circumstances, maximum-likelihood and parsimony methods can be shown to be formally equivalent¹⁶.

For traits that naturally vary along a continuous scale, constant-variance random-walk (sometimes called brownian motion) models are the analogue to the Markov-transition model. In the conventional random-walk model, traits evolve in each instant of 'time' dt with a mean change of zero and unknown and constant variance, σ^2 . Time may be chronological or some other unit of divergence, such as genetic distance. The evolutionary process is

Box 1

Statistical terms and models

Generalized least squares (GLS)	A statistical method suitable for analysing data that may not be statistically independent or may not have the same expected variances ¹⁷ . It is well suited to phylogenetic applications, owing to the expected similarity amongst species associated with phylogenetic relatedness, and to the different variances arising from differing total path lengths from the root to species. Neutral-drift and directional GLS models both use these features, but the directional model can also detect trends of trait evolution.
Independent contrasts	A method for analysing data from phylogenies that uses the phylogeny to identify a set of mutually independent comparisons between pairs of species, pairs of nodes, or a node and a species (see refs 54, 58).
Likelihood	An amount proportional to the probability of observing the data, given some model ³ . In an evolutionary context, the model often characterizes an aspect of trait evolution.
Likelihood ratio	A way of statistically testing for a difference between two likelihoods (instantiations of models).
Maximum likelihood	A set of techniques for choosing the parameters of a statistical model in such a way as to provide the most probable description of the observed data, given the model.
Markov process	A process in which the probability that a trait takes some value depends only on the value of the trait in the previous unit of time — the Markov no-memory property. For discrete characters, like nucleotides or amino acids, the probability of being in state i at time t depends only on what the state was at time $t - 1$.
Maximum parsimony	In a phylogenetic context, a set of techniques for reconstructing phylogenies or ancestral states on phylogenies that prefers solutions requiring the least amount of change.
Random walk	A process in which the state of a variable at time t is given by its starting point (time 0) plus the sum of all the random changes to the variable from time 0 to time t . In the neutral-drift model, changes are randomly sampled from a distribution with a mean of zero and a fixed variance. In a directional drift model, the mean of the distribution of changes is different from zero.
Squared-change parsimony	A method for reconstructing ancestral character states that minimizes the sum of squared changes along the branches of the phylogenetic tree. It produces the same estimates as methods used in independent contrasts.

presumed to unfold independently at each instant of time and along each of the branches of the phylogeny. The expected variance of a given species' trait value is then $\sigma^2 t$, where t records the total path length (time or distance) from the root to that species. Trait values of species or lineages that have diverged more from the root are expected to have larger variances and so are less reliable observations. Closely related species will tend to have similar trait values, even under a random walk, as they share most of their evolutionary history. A generalized least squares (GLS) approach^{2,8,17,18} provides a natural framework in which to represent these features that arise from phylogenetic associations, while simultaneously estimating the variance parameter of trait evolution.

The standard GLS model, by presuming that traits evolve according to an unbiased random walk (neutral drift), cannot detect any directional trends of trait evolution along the branches of the tree. Historical trends such as a phyletic increase in size (for example, Cope's law) will be masked. Consequently, this model always estimates the ancestral state at the root of the phylogeny as falling somewhere within the range of observed values in the species data. Ancestral states obtained from squared-change parsimony or from the popular independent-contrasts approaches for comparative studies are equivalent to those obtained from the constant-variance random-walk model⁸. A general treatment for reconstructing ancestral states of continuously varying characters can be found in ref. 8.

Recent directional GLS models^{2,18} for continuous traits can detect historical trends of trait evolution. These models examine the correlation between the species' trait values and the total phylogenetic distance or path length from the root of the tree. If a

directional trend exists, species that have diverged more from the root will also tend to have changed more in a given direction, that is, they will be larger or mature earlier, for example. Under these circumstances the directional model will reconstruct ancestral states more accurately and, importantly, it can use the trend to reconstruct the character state at the root of the tree to lie outside of the range of observed values in the data.

Ribonuclease evolution. Maximum-parsimony reconstructions of ancient artiodactyl ribonucleases¹⁹—an enzyme involved in foregut digestion—assign glycine as the ancestral state at a key amino-acid position, with a transition to aspartic acid about 40 million years ago (Fig. 2). This transition may correspond to the adoption of true ruminant digestion¹⁹. Maximum-likelihood reanalyses of the same data⁷ reveal high rates of character evolution at this site: the estimated 'half-life' for replacement of glycine by aspartic acid is approximately 42 million years (Myr) and for aspartic acid by glycine is 68 Myr (ref. 7), over a tree length spanning 450 Myr. The relatively short half-lives arise because the phylogenetic distribution of traits (Fig. 2) imply that there were at least three historical changes of the amino acid in these artiodactyls. Likelihood reconstructs the ancestral state as aspartic acid, opposite to the inference derived from parsimony (Fig. 2) and implies at least four, as opposed to three, evolutionary transitions on the tree.

Why is there a difference? The likelihood analysis detects that at least two of the four aspartic acid-to-glycine replacements that its reconstructions minimally imply occur in long branches (those leading to camels and cows). These changes are not improbable owing to the high rates of evolution. The parsimony analysis, by ignoring the branch-length information, probably underestimates the amount of change in these longer branches. Parsimony therefore prefers the solution of three events to four, even though these events are reconstructed to occur in comparatively short branches, and require one reversal (an aspartic acid to glycine) transition. Likelihood considers these changes to be relatively improbable. If the branches were all set to the same length, parsimony and likelihood would return the same answers for these data. Maximum likelihood seeks the most probable explanation of the observed data (including

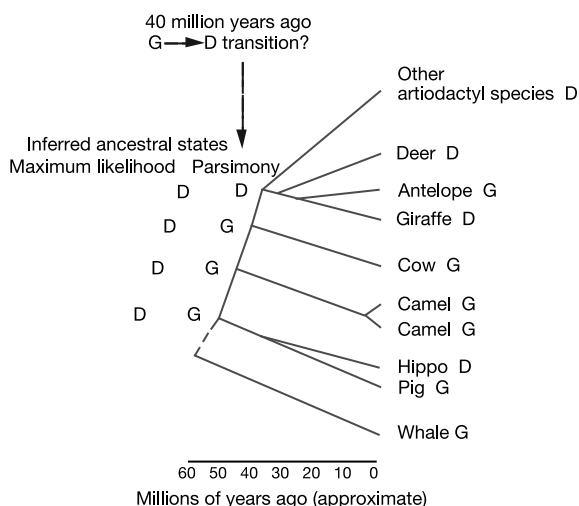


Figure 2 A phylogeny of the artiodactyls with approximate branch lengths, adapted from ref. 19. More recent phylogenies place the hippopotamus as the sister group to whales²⁷, but this difference does not qualitatively affect the results reported here. The G (glycine) and D (aspartic acid) indicate the species values, and, separately, the ancestral states derived from parsimony¹⁹ and maximum likelihood⁷. Jermann *et al.*¹⁹ postulate a transition from G to D at position 38 of the ribonuclease gene, corresponding to the adoption of true ruminant digestion in artiodactyls. Maximum-likelihood methods favour D as the ancestral state (see text). Whales are included to show a putative outgroup. The line to the whales is dotted and the root values are not recorded because Jermann *et al.*¹⁹ and Schluter⁷ considered only the artiodactyls.

the tree and branch lengths), given the model of evolution, not necessarily the solution with the fewest events.

The likelihood analysis raises doubts about the suitability of parsimony for these data. It also raises the fundamental question of whether parsimony is the best general model for trait evolution. The Markov-transition model in a likelihood framework has an advantage in that it takes into account the length of branches in the tree and can adjust to low or high rates of change. The ribonucleases may not be an isolated case: parallel and convergent evolution at the molecular level may be more common than once believed^{20–24}, and examples of high rates of morphological trait evolution are easy to find¹¹.

The common ancestor to life. The common ancestor to all cellular life may have arisen more than 3.8 billion years ago²⁵ when the Earth's environment was hot, unstable and wracked by volcanoes. Phylogenetic trees of life typically reveal that the extant hyperthermophilic bacteria and archaeal species, which inhabit environments of extreme temperatures, have some of the deepest and oldest branches^{26–28} (Fig. 3), and it is consequently a widely endorsed textbook view that the common ancestor of life was adapted to hot conditions.

The proportion of all nucleotides that are either guanine or cytosine (the G+C content) of ribosomal RNA is a reliable indicator of the environmental temperature of an organism²⁹, so an estimate of the G+C content of the root of the tree of life provides evidence for the environmental conditions that prevailed when the common ancestor to life arose. The technical difficulty in estimating the G+C content of the common ancestor to life lies in a limitation of the standard Markov model of gene-sequence evolution as applied to estimating phylogenetic trees. These models are classified as homogeneous and stationary, that is, they presume for computational purposes that all lineages have the same frequencies of nucleotides (often taken to be the relative frequencies of A, G, C, and T or U in the data), and that all lineages are in equilibrium in the sense that these relative frequencies have not changed over time. Observed differences in base composition among lineages are treated as chance variation arising from neutral drift. A consequence of these assumptions is that homogeneous and stationary Markov models are constrained always to assign the ancestral nucleotide frequencies within the range observed in the data, typically as the relative frequencies observed in the extant species.

In an important development, Galtier *et al.* (ref. 29) suggest a non-homogeneous, non-equilibrium Markov model of gene-sequence evolution. The model allows lineages to evolve lineage-specific relative nucleotide frequencies, thereby exploiting the observation that lineages separated for billions of years can have

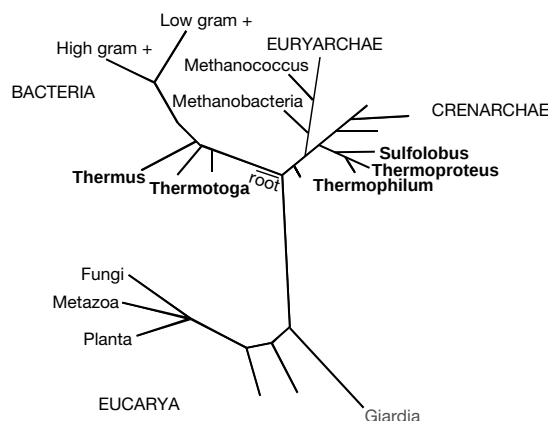


Figure 3 A tree of life showing the approximate relationships among the bacteria, archaea (euryarchaea and crenarchaea) and eukaryotes, adapted from ref. 27. Extant hyperthermophilic organisms are in bold and are close to the root of the tree of life.

Table 1 G+C nucleotide content for the four main groups in the tree of life

Group	All species (LSU)	G+C nucleotide content	
		Extreme sequences	
		SSU	LSU
Eucarya	50.98%	71.0, 70.9	65.7, 62.8
Crenarchae	59.63%	65.4, 65.2	60.4, 59.7
Euryarchae	55.08%	65.1, 65.0	58.7, 58.4
Bacteria	52.78%	62.4, 63.2	58.3, 57.4
Estimated ancestral state (common ancestor to life)	54.0%	57.3%	55.5%
Typical hyper-thermophile		≥60%	

Extreme sequences represent the two species with the highest G+C content within each group (from ref. 29). Sequences are from short-subunit (SSU) and long-subunit (LSU) ribosomal RNA.

widely divergent G+C contents. If the nucleotide frequencies diverge between lineages systematically with underlying genetic divergence, then a directional trend of evolution is suggested, rather than neutral drift. The model can detect such directional trends, and so can potentially estimate ancestral states that lie outside the range of features observed in the species data.

Galtier *et al.*'s model estimates the ancestral G+C content of both short- and long-subunit ribosomal RNA at an intermediate value of 52–54%, with narrow confidence limits, when applied to sequences from the euryarchae, eucarya, crenarchaea and bacteria, including thermophilic and non-thermophilic species (Table 1). Analysis of only those species with the highest G+C content from each of the four principal groups shows that these results are not simply some average of the observed data: extraordinarily, the estimated ancestral sequences for this subset have lower G+C content than any of the extreme sequences. Computer analyses confirm the reliability of the estimates and the capability of the model to reconstruct the ancestral sequence correctly in simulated data²⁹.

Differences in the G+C contents of ribosomal sequences in lineages that diverged perhaps more than three billion years ago preserve the trace of an ancient and long-term directional trend of evolution. The non-equilibrium statistical model recovers these traces, enabling it, when evolution is traced backwards in time, to estimate the ancestral G+C content at a value lower than suggested from previous analyses. The significance of Galtier's *et al.*'s results²⁹ is that the G+C content in the ranges they estimate is not compatible with the high G+C content (typically 60 or more) that would indicate a thermophilic common ancestor to life.

Extant hyper-thermophilic species, rather than being ancestral, may be derived descendants of a common ancestor to life that arose in environments of moderate temperature: their deep-rooting phylogenetic position is not a reliable indicator of the ancestral G+C content. Reminiscent of the ribonuclease genes discussed above, hyper-thermophilia may have evolved independently several times. The reconstructed ancestral states do not unequivocally rule out a thermophilic origin of life, but no evidence for it can be claimed from the RNA sequences most likely to harbour that evidence.

Influenza evolution. Directional models can prove crucial for correctly estimating the ancestral state of a continuous trait. Figure 4 shows the cactus-like structure of phylogenetic trees of the haemagglutinin and non-structural 1 (NS1) influenza genes, based on samples drawn at known times over a number of years³⁰. Branches that end before the present represent extinct lineages of the influenza virus. The plots record the number of nucleotide substitutions to each gene, as measured from the root of the tree, against the year in which the sample was obtained, and reveal striking clock-like regularity.

The dates predicted to fall at the roots of these two trees correspond to the occurrence of the common ancestors to the haemagglutinin and NS1 genes, that is, to the genes that would have been ancestral to all of the lineages in the respective trees. The predicted date of the root must be earlier than any of the obser-

vations but, for both genes, the drift GLS model predicts that the common ancestor arose after at least one of the known sampling times in the data set (haemagglutinin, 1968.2 ± 0.4; NS1, 1939.1 ± 5.7; asterisks on the time axis of Fig. 4).

What has gone wrong? The directional GLS model (dashed regression lines) identifies that time and molecular divergence strongly covary in these data. By rewinding time down the branches leading to the root, the directional model correctly places the time of the common ancestor outside the range of observed values (the point where the dashed regression lines cross the time axis: haemagglutinin, 1968.8 ± 0.1; NS1, 1931.20 ± 2.3; Fig. 4). The slopes of the directional regression indicate that there were about 6.5 and 1.7 substitutions each year, respectively, for haemagglutinin and the NS1 gene.

The solid regression lines (Fig. 4) are derived from a simple regression across the observed data. Although here they are similar to the GLS regressions, they are incorrect as they fail to take account of the similarities (non-independence) among lineages that arise from phylogenetic associations, and by virtue of giving equal weight to all observations despite large differences in their expected variances. The NS1 regression line, when extrapolated, crosses the time axis at a point after at least two of the observations in the data set.

Directional trends, such as those identified for influenza and in the analysis of G+C content, can reveal fundamental trends in the history of trait evolution that must be accounted for if the past is to be reconstructed accurately. This is true whether the traits under investigation are time or some morphological, life-history, behavioural, genetic or other feature. Little is known about such trends outside careful palaeontological research (see refs 31, 32 for exceptions), but appropriate statistical models can identify them and hence return provocative new answers to old questions.

Timings of evolutionary events

If d is the amount of gene-sequence divergence between two species, and r is the rate of molecular evolution, then their time of divergence is given by $T_{\text{past}} = d/2r$, where the 2 accounts for the fact that the genetic divergence has accumulated independently in two branches. This is the widely used molecular clock. The rate of evolution r is obtained by calibrating a given amount of divergence from the date of a fossil or geological event. However, r may vary from lineage to lineage and from time to time. This rate heterogeneity greatly limits the accuracy and credibility of molecular-clock studies³³.

New statistical models are limiting the damage of this critique. A maximum-likelihood approach based on the 'quartet' method can accommodate rate heterogeneity from two independently calibrated clocks to estimate times of divergence^{34,35}. The method is applied to two pairs of species (Fig. 5) for which fossil ancestors are known. A calibrated molecular clock is derived for each pair by using the known dates and genetic divergences between the pairs. The two clocks are then used simultaneously to estimate the most likely date of the common ancestor to the four species. Computer simulations show that the method returns good estimates of times of divergence^{34,35}.

By combining the information from two independently calibrated molecular clocks, the quartet method can accommodate rate heterogeneity between the two pairs of species. However, the method assumes that a constant molecular clock operates within each pair of species in the quartet. This may not be true for any given data set, and so rate constancy within the pairs is tested by fitting a standard Markov model of sequence evolution to the unrooted quartet. The quartet is unrooted at this step because the date at the root is unknown, so the two oldest branches (Fig. 5) are joined. Rates of evolution are allowed to vary freely in each of the five branches of the unrooted quartet. Rate constancy is rejected if this five-rate model fits the data better than does the constrained (two-

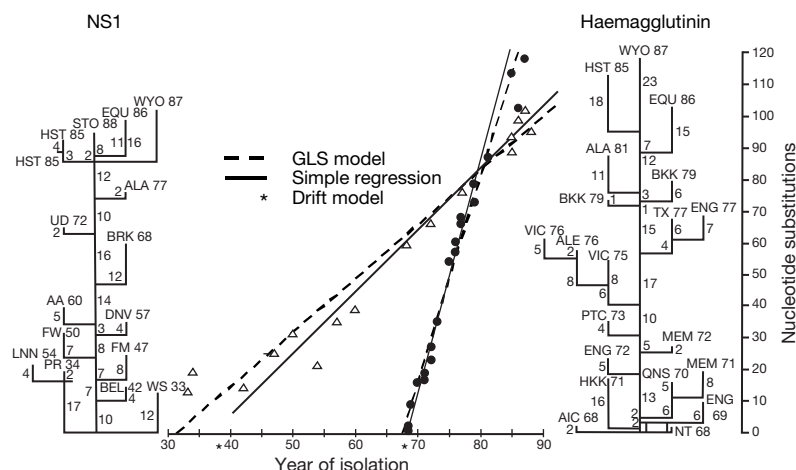


Figure 4 Peculiar cactus-like shape of two gene trees from data on the influenza virus (modified from ref. 30). Left, non-structural 1 (NS1) gene. Right, haemagglutinin gene. Branches that terminate before the present (top) are extinct lineages; sampling times are indicated in the lineage name (for example, ALA77 was obtained in 1977). The length of the vertical lines in the gene trees is proportional to the number of nucleotide substitutions (indicated by the number adjacent to the vertical line). The total number of substitutions inferred to have occurred since the root of the tree is shown in the scale on the right. The graph records the number of substitutions from the root to the tip against the year of sampling; triangles, NS1 gene; circles, haemagglutinin. Three statistical analyses are shown. Asterisks along the time axis indicate inferred root times, as derived from the neutral-drift GLS model; this model predicts the root but ignores the information linking

time to genetic divergence. The dashed regression line is derived from the directional GLS model; this model predicts the root and therefore extends to the time axis, signifying the time of the point of zero genetic divergence. Solid regression lines are derived from a simple across-species analysis; the simple regression does not extend to the time axis (and thereby predict the value at the root) because this method fits the line only to the observed data; extrapolating beyond these data is not valid. The directional-GLS and simple regressions differ because the latter weights all observations equally, whereas the GLS models give more weight to observations closer to the root as they are expected to have smaller variances (see text). The simple regression also fails to account for the non-independence among successive data points that derives from counting nucleotide substitutions up the backbone of the tree repeatedly.

rate) model, in which case the gene sequences are excluded as unreliable clocks. Only when rate constancy is supported are sequence data used to estimate molecular clocks, which in turn provide an estimate of divergence time.

The Cambrian explosion? The quartet method returns extraordinary results that challenge conventional wisdom about the dates of the major adaptive radiations of animal phyla associated with the Cambrian explosion. According to the Cambrian-explosion hypothesis, the major phyla and even classes of the animal kingdom emerged suddenly in a rapid evolutionary radiation beginning about 565 Myr ago near the start of the Cambrian era³⁶. Representatives of nearly all animal phyla that fossilize are found in the Cambrian rocks. The earliest undisputed metazoan fossils are dated to about 600 Myr³⁷, reinforcing the view of the Cambrian as a sort of 'Big Bang' of animal evolution³⁸.

Quartet-model estimates derived independently from a large number of genes indicate that the ancient echinoderm–vertebrate and protostome–deuterostome lineages of multicellular animals diverged long before the Cambrian³⁵. Excluding genes showing rate heterogeneity within pairs, most divergences are estimated at 1,000 Myr or more, and none occurs more recently than 680 Myr; 95% confidence intervals exclude both the Cambrian origin and the earlier fossil dates of around 600 Myr.

Other recent molecular estimates of the divergence of animal phyla³⁹, derived from genes tested for rate constancy, broadly support an age of 1,000 Myr. Suggested dates of 704 Myr and 600–670 Myr from the reanalysis⁴⁰ of data³⁹ may be based on underestimates of the rate of the molecular clock³⁵. The pre-Cambrian has been described as more of a 'phylogenetic fuse'³⁵ than a period leading up to a Big Bang. New palaeontological findings lend support to the earlier estimates derived from the molecular data. Controversial fossil evidence of worms 1,000 Myr old has been discounted, but some consensus is emerging for a figure of 680 Myr⁴¹, pre-dating the Cambrian by 120 Myr—a long time in the life of an 'explosion'.

Effects of the K–T extinction. The quartet method returns similarly starting results when applied to the effects of the K–T extinction

event. The K–T boundary describes a time approximately 65 Myr ago when the dinosaurs became extinct, possibly as a result of a cataclysmic collision between Earth and an asteroid. Geological and fossil evidence indicate that up to 70% of mammalian and avian taxa were lost⁴², and the classical view based on fossil evidence is that the major avian and mammalian orders diverged rapidly in adaptive radiations thereafter. The present diversity of mammalian and avian life on Earth is suggested to be largely an accidental by-product of those lineages lucky enough to survive the K–T extinction⁴².

Estimates derived from the quartet method tell a different story, however. At least 22 diverse avian lineages probably survived the K–T extinction⁴³. Molecular-clock estimates of the timings of the ordinal diversification of birds and mammals, based on genes tested for rate constancy, substantially precede the K–T event, falling 90 Myr to 113 Myr⁴⁴. Intriguingly, these dates for bird and mammal diversification coincide instead with the break-up of large continental land masses⁴⁴.

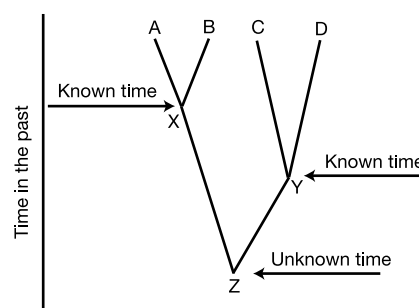


Figure 5 The quartet method simultaneously considers two pairs of species with known times of divergence based on fossils. The genetic distance between each pair is calculated and the molecular clock calibrated from the divergence time, yielding rate estimates μ_x and μ_y . The unknown branch lengths XZ and YZ can then be found from $XZ = \mu_x(t_z - t_x)$ and $YZ = \mu_y(t_z - t_y)$, where t_z is chosen in a standard Markov model of gene-sequence evolution to maximize the likelihood of observing the known times and rates (after Fig. 1 in refs. 34, 35).

Re-evaluating animal diversification. So divergent and provocative are these new findings pertaining to the Cambrian and K–T extinction that they force a re-evaluation of both molecular and fossil data in an effort to understand the evolution of animal diversity millions of years ago. Is it possible that the molecular clocks are wrong? If rates of gene-sequence evolution were higher during the Cambrian period, clocks calibrated from more recent (slower) periods would overestimate the age of the divergence of ancestral lineages. However, Wray *et al.* (ref. 39) provide a detailed discussion of why it is unlikely that undetected rate acceleration during the Cambrian would alter the conclusions of the molecular-clock studies. The earlier dates derived from the new statistical models do not deny the existence of a rich diversity of forms in the Cambrian, but they do suggest that the basic animal body plans emerged long before the Cambrian. Similarly, the K–T events were momentous but the diversity of avian and mammal forms may owe more to adaptation to new environments arising from gradual tectonic, rather than sudden cataclysmic, climatic events.

Molecular-clock studies are providing a rich and independent set of results for anthropologists, biologists and palaeontologists to debate. The date of the domestication of dogs has recently been set back to approximately 135,000 years ago⁴⁵, far earlier than the 14,000 years ago favoured by archaeologists, and may have occurred several times independently in different regions of the world. The long-awaited amplification of ancient DNA from the Neanderthal type specimen indicates that the lineages leading to Neanderthals and modern humans diverged 500,000 years ago⁴⁶, weighing against speculation and some fictional accounts of gene flow between these groups.

The tempo of evolution

Do traits evolve in fits and starts, or smoothly and gradually? Darwin's preference for gradualism is evident in his remark that "as natural selection acts solely by accumulating slight, successive, favourable variations, it can produce no great or sudden modifications"⁴⁷, although some interpret him differently⁴⁸. Proponents of punctuated equilibria⁴⁹ contend that the fossil record suggests a different view: that most evolution occurs during relatively short periods of rapid change that are interspersed within far longer periods of stasis, during which time traits are in some form of equilibrium.

The challenge of punctuated equilibria to gradualists is that it begins to unhook the evolution of traits from the regular tug of natural selection. Theorists have responded imaginatively⁵⁰. The sometimes intemperate debate between the two camps has continued for over 25 years.

Critics charge that the fossil record can never provide sufficient resolution to answer questions of the tempo of trait evolution convincingly (see, for example, ref. 51), whereas others emphasize the completeness of the record for some groups⁵². It is little appreciated that a continuum of trait evolution from gradual to punctuational change can be detected on phylogenies, given appropriate scaling of branch lengths. Key questions of gradual versus punctuational evolution of traits can be tested, and over a far wider range of both taxa and traits than is available in the fossil record.

Maximum-likelihood methods in conjunction with the models for discrete and continuously varying traits can estimate scaling parameters to characterize the tempo of trait evolution. The parameter called κ defines for discrete traits²⁶ the relationship between the lengths of individual branches and the probability that a character changes state. For continuously varying characters, the parameter δ scales both the shared phylogenetic path lengths between related species and the total path lengths from the root of the phylogeny to the tips (defined as continuous-variables κ in ref. 6).

These parameters discriminate amongst several modes of trait evolution. Direct gradualism describes traits that change linearly with branch length (discrete traits, $\kappa = 1.0$), or with shared and

total path length (continuous traits, $\delta = 1.0$). Scaled gradualism arises when the relationship of trait evolution to branch or path length is nonlinear; κ or $\delta > 1.0$ implies that traits change proportionately more in longer branches, or that longer paths contribute more to trait evolution, as would be true if later (that is, more recent) evolution has contributed more than earlier events; κ or $\delta < 1.0$ indicates traits changing rapidly at first then remaining stable, as might occur in adaptive radiations. Decoupled trait evolution refers to traits evolving by amounts independent of the lengths of the branches or path lengths in the tree (κ or $\delta = 0.0$). Scaled gradualism with κ or $\delta \ll 1.0$ and decoupled trait evolution may be consistent with punctuated equilibria. An alternative to these scaled models allows the rate of trait evolution to vary independently in every branch of the tree⁵³.

The two influenza genes (Fig. 4), although apparently clock-like in their rates of evolution, yield different trends when δ is used to scale sampling time against number of substitutions. The maximum-likelihood estimate of δ for haemagglutinin is not different from 1.0 ($\hat{\delta} = 1.06$, 95% confidence intervals = 0.89–1.24), indicating that this gene is indeed evolving chronometrically. In contrast, the clock-like behaviour of the NS1 gene hides a more complex tempo. The maximum-likelihood estimate of δ is 0.57 (95% confidence interval = 0.22–0.93), indicating that the gene evolves rapidly in shorter paths (temporally earlier lineages in this example), then slows in the later surviving lineages. In other words, more recent viruses have changed less year to year than did earlier ones, indicating that the gene may have reached some sort of fitness plateau. Substitutions to the ribonuclease gene (Fig. 2) do not depart from direct gradualism ($\hat{\kappa} = 0.98$; 95% confidence intervals = 0.52–1.25), providing further evidence that ancestral reconstructions that take account of branch lengths are preferable for these data. As Fig. 4 shows, trends such as those reported here may not be apparent from an inspection of data, but emerge when the data are subjected to the right statistical analysis.

Correlated evolution

The comparative method is one of evolutionary biology's most enduring traditions for testing hypotheses of adaptation⁵⁴. Comparative methods seek evidence for adaptation in the patterns of correlated trait evolution across contemporary species, that is, in how characteristics of organisms (such as size, shape, life history and behaviour) evolve together. The correlations derived from a group of contemporary organisms can also be shown to have an historical interpretation: under some models of evolution, they reflect the correlation that would have held in the ancestral species as the same traits evolved along what would become the branches of phylogenetic trees⁵⁵.

Most investigators now realize that statistical analyses of species must take into account the fact that closely related species tend to be more similar than distantly related ones, and thus that species cannot be considered as statistically independent units of observation⁵⁴. Conventional comparative methods for discrete variables use the phylogeny to reconstruct by parsimony the probable ancestral character states, and then calculate statistics conditional upon those reconstructions^{56,57}. The 'independent contrasts' approach^{54,55,58–62} has come to dominate comparative analyses of continuously varying characters. These methods use the phylogeny in combination with a neutral-drift model of trait evolution to identify a set of statistically independent comparisons among species.

New maximum-likelihood models of correlated trait evolution for discrete⁶ and continuously varying characters^{2,18} offer hypothesis tests not available to the earlier methods, and avoid some of the difficulties of those approaches. Correlated evolution of discrete binary characters is modelled in a maximum-likelihood framework by the same Markov-process model as described for reconstructing ancestral character states of discrete characters. Evidence for corre-

lated change in two characters is gathered by comparing the likelihood of two binary characters, each allowed to evolve independently on the tree, to the likelihood obtained when the same characters are modelled as if they are evolving in a correlated fashion^{2,6}.

Statistical tests for the dominant direction of change (for example, do 'forward' transitions predominate over 'backward' ones) and the temporal order of changes (for example, does character *y* change before or after character *x*) are possible by comparing appropriate likelihoods⁶. The temporal-order test detects traces of the probable sequence of evolution from the ancestral states of two characters to the derived states of both (see, for example, ref. 63).

Maximum-likelihood models for investigating correlations amongst continuously varying characters use the same GLS framework as described for modelling trait evolution in the influenza virus². The models return estimates of the correlations and regressions among two or more variables while controlling for phylogenetic associations. Trait evolution along the branches of the phylogenetic tree can be modelled in the GLS framework by the neutral drift or other models, such as the model of directional change.

Under the neutral-drift model, the GLS and independent-contrasts techniques^{58,59} return the same correlation. This 'drift correlation' includes variance and covariance in the traits that arise from at least three sources: neutral drift, any directional or other non-random trends, and independent correlated trait evolution. By comparison, in the 'directional correlation', the variance arising from neutral-drift and directional trends is removed. The directional-model correlation therefore estimates solely the independent correlated trait evolution, and will differ from the drift correlation whenever there are directional trends of trait evolution in the data.

A key advantage of the GLS approach, however, lies in its ability to scale phylogenetic path lengths in response to patterns in the data². Several recent studies show that independent-contrasts methods perform worse than simple non-phylogenetic analyses in some circumstances (refs 64–66; P. H. Harvey & A. Rambaut, unpublished data). In one of these⁶⁵, traits are modelled to evolve by a larger amount per speciation event early on than later in the phylogeny, and trait values are not necessarily very similar between pairs of closely related species. In short, the variance of evolutionary change is not constant, so the neutral-drift model that the independent-contrast method presumes (see the section Reconstructing ancestral states) does not accurately reflect trait evolution.

Using the GLS approach, the δ scaling parameter, described in conjunction with characterizing the tempo of evolution, would detect early and rapid trait evolution as $\delta < 1.0$ and rescale the path lengths accordingly. A second scaling parameter, denoted λ , detects whether the shared evolutionary histories as specified by the phylogeny produce the patterns of similarity observed in the data. Values of $\lambda < 1.0$ correspond to traits being less similar amongst species than expected from their phylogenetic relationships; $\lambda > 1.0$ suggests the reverse. We might expect $\lambda < 1.0$ for the model of ref. 65. Signatures of trait evolution relevant to the suitability of a comparative method reside in the data. The GLS model, by detecting these signatures, is expected to perform reasonably well, even under circumstances averse to other methods.

The evolution of lactose tolerance. The enzyme lactase confers an ability to digest milk. Human infants can digest milk, but most adults cannot (the widespread tolerance to lactose among some European groups is an exception). The dominant view is that adult lactose tolerance in humans is an adaptation to reduced exposure to the Sun⁶⁷. Both the Sun and the enzyme lactase promote calcium absorption. This hypothesis accounts for the lactose tolerance of some northerly dwelling cultural groups, such as the reindeer-herding Lapps, and the prevalence of lactose tolerance in some European groups.

Table 2 Slope of the regressions of brain volume and basal metabolic rate against body size in mammals

Variable	Species	Correlation	GLS regression slope (95% confidence interval)	Theoretical expectation
Brain volume	23	0.96	0.59 (0.52–0.67)	0.67–0.75
Basal metabolic rate	15	0.95	0.72 (0.60–0.84)	0.75

An alternative to the latitude theory is that adult tolerance to lactose is advantageous in cultures that keep animals for milk. If milk forms a significant portion of the diet, selection pressures on adults to develop the ability to digest it could be strong.

Holden and Mace⁶⁸ applied the Markov-process model⁶ to a phylogeny of human cultural groups to show that adult lactose tolerance has arisen independently—possibly by selection for rare alleles—up to three times in cultures that keep animals for milk, but not in non-dairying cultures. The latitudinal effect disappeared when the phylogeny was used to take account of non-independence amongst cultural groups. Likelihood-ratio tests of the temporal order of gene versus cultural change reveal the probable course of the evolution of lactose tolerance. The ancestral condition of non-dairying and no adult lactose tolerance was first replaced by dairying perhaps as early as 6,000 to 8,000 years ago, which then favoured tolerance to lactose. The data do not support the alternative scenario that human groups with adult lactose tolerance were more likely to adopt dairying.

The phylogenetic statistical approach is able to untangle the latitudinal and dairying effects in a way that previous analyses could not, and could detect evidence to confirm the important causal order of effects: cultural practices of herding societies appear to have selected for a genetic adaptation.

Mammalian brain-size evolution. Theories of the evolution of mammalian brain size link brain volume to body mass, either through body surface area or through basal metabolic rate^{69,70}. The surface-area hypothesis requires a slope of 2/3 between brain volume and body mass; the metabolic-rate hypothesis predicts that the slope of brain volume against body mass will be 3/4, owing to the relationship of basal metabolic rate to body size. Empirically derived slopes generally lie somewhere in the 0.67–0.75 range⁷¹. New molecular phylogenies of the mammals derived from whole mitochondrial DNA (mtDNA) sequences⁷² afford tests of the surface-area and metabolic-rate hypotheses using the features of the GLS models.

Before calculating a correlation, the maximum-likelihood values of the path-length scaling parameter δ and the phylogeny scaling factor λ were estimated: both overlap 1.0 ($\hat{\delta} = 0.96$ (0.078–2.79) and $\hat{\lambda} = 1.0$ (0.94–1.0)), indicating that brain-size evolution is proportional to the branch lengths of this mtDNA phylogeny, and justifying the use of the unscaled phylogeny to correct for non-independence. The GLS model returns a slope relating brain volume to body mass that is lower than the expected 2/3 to 3/4, despite a very strong correlation between the two traits (Table 2). Even in these comparatively small samples, the 95% confidence intervals exclude 3/4 and nearly exclude 2/3. The same GLS applied to a composite primate phylogeny⁷³ of 59 species yields a slope of 0.48 (95% confidence intervals = 0.39–0.57) between brain and body size².

Are the regression slopes peculiar to the model or phylogeny? Applying the same model to estimate the slope of basal metabolic rate against body size, again using the mtDNA phylogeny, gives an empirical result (Table 2) that is very close to the expected value of 3/4 (ref. 70). In the light of these results, the surface-area and metabolic-rate theories for mammalian brain size would seem increasingly difficult to support. Linking brain adaptations to specific environmental demands^{71,74} may offer an alternative way of thinking about brain evolution.

Conclusions

The development of statistical modelling techniques for analysing trait evolution on phylogenies is still in its early years but, as the studies reported here show, they are already returning results that question long-standing views of the history of life and the patterns of adaptation. They also raise questions about how we should think about evolution. Is it smooth and gradual, or punctual? Is parsimony, so long the unquestioned null model of trait evolution, really appropriate, or should it be replaced by models that explicitly consider the lengths of the branches of phylogenetic trees? Future developments^{75,76} may make it feasible to estimate the historical processes of evolution simultaneously with estimation of the phylogenetic tree.

The growth of statistical inference techniques brings a new and independent point of view to debates about the history of life. There are bound to be many factional and territorial disputes as their use grows, however. As predominantly visual and tactile primates, we may find it far harder to accept statistical inferences than that which we can see and touch. But as in any historical discipline, new ideas will be judged by how well they explain the existing facts. Statistical techniques for inferring the history and pattern of evolution will prove invaluable for looking into the past in ways, and for kinds of traits, that are out of reach to other approaches. □

Received 21 May; accepted 7 September 1999.

- Hillis, D. M., Moritz, C. & Mable, B. K. *Molecular Systematics*, 2nd edn (Sinauer, Sunderland, Massachusetts, 1996).
- Pagel, M. Inferring evolutionary processes from phylogenies. *Zool. Scripta* **26**, 331–348 (1997).
- Edwards, A. W. F. *Likelihood* (Cambridge Univ. Press, Cambridge, 1972).
- Golding, G. B. & Dean, A. M. The structural basis of molecular adaptation. *Mol. Biol. Evol.* **15**, 355–369 (1998).
- Ivics, Z., Hackett, P. B., Plasterk, R. H. & Izsvak, Z. Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell* **91**, 501–510 (1997).
- Pagel, M. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond. B* **255**, 37–45 (1994).
- Schluter, D. Uncertainty in ancient phylogenies. *Nature* **377**, 108–109 (1995).
- Schluter, D., Price, T., Mooers, A. Ø. & Ludwig, D. Likelihood of ancestor states in adaptive radiation. *Evolution* **51**, 1699–1711 (1997).
- Yang, Z., Kumar, S. & Nei, M. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* **141**, 1641–1650 (1995).
- Koshi, J. M. & Goldstein, R. A. Probabilistic reconstruction of ancestral protein sequences. *J. Mol. Evol.* **42**, 313–320 (1996).
- Mooers, A. Ø. & Schluter, D. Support for one and two rate models of discrete trait evolution. *Syst. Biol.* **48**, 623–633 (1999).
- Pagel, M. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Syst. Biol.* **48**, 612–622 (1999).
- Hillis, D. M., Bull, J. J., White, M. E., Badgett, M. R. & Molineux, I. J. Experimental phylogenetics: generation of a known phylogeny. *Science* **255**, 589–592 (1992).
- Collins, T. M., Wimberger, P. H. & Naylor, G. J. P. Compositional bias, character-state bias, and character-state reconstruction using parsimony. *Syst. Biol.* **43**, 482–496 (1994).
- Maddison, D. R. Phylogenetic methods for inferring the evolutionary history and process of change in discretely valued characters. *Annu. Rev. Entomol.* **39**, 267–292 (1994).
- Tuffey, C. & Steel, M. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bull. Math. Biol.* **59**, 581–607 (1997).
- Johnston, J. *Econometric Methods* (McGraw Hill, New York, 1963).
- Martins, E. P. & Hansen, T. F. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am. Nat.* **149**, 646–667 (1997).
- Jermann, T. M., Opitz, J. G., Stackhouse, J. & Benner, S. A. Reconstructing the evolutionary history of the aridodactyl ribonuclease superfamily. *Nature* **374**, 57–59 (1995).
- Li, W.-H. *Molecular Evolution* (Sinauer, Sunderland, Massachusetts, 1997).
- Yeager, M. & Hughes, A. L. Evolution of the mammalian MHC: natural selection, recombination, and convergent evolution. *Immunol. Rev.* **167**, 45–58 (1999).
- Bull, J. J. *et al.* Exceptional convergent evolution in a virus. *Genetics* **147**, 1497–1507 (1997).
- Roux, K. H. *et al.* Structural analysis of the nurse shark (new) antigen receptor (NAR): Molecular convergence of NAR and unusual mammalian immunoglobulins. *Proc. Natl Acad. Sci. USA* **95**, 11804–11809 (1998).
- Crandall, K. A., Kelsey, C. R., Imamichi, H., Lane, H. C. & Salzman, N. P. Parallel evolution of drug resistance in HIV: Failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *Mol. Biol. Evol.* **16**, 372–382 (1999).
- Mojzsis, S. J. *et al.* Evidence for life on earth before 3,800 million years ago. *Nature* **384**, 55–59 (1996).
- Woese, C. R. Bacterial evolution. *Microbiol. Rev.* **51**, 221–271 (1987).
- Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740 (1997).
- Brown, J. R. & Doolittle, W. F. Archaea and the prokaryote-to-eukaryote transition. *Microbiol. Mol. Biol. Rev.* **61**, 456–504 (1997).
- Galtier, N., Tourasse, N. & Gouy, M. A nonhyperthermophilic common ancestor to extant life forms. *Science* **283**, 220–221 (1999).
- Fitch, W. M., Leiter, J. M. E., Li, X. & Palese, P. Positive Darwinian evolution in human influenza A viruses. *Proc. Natl Acad. Sci. USA* **88**, 4270–4274 (1991).
- Omland, K. Character congruence between a molecular and a morphological phylogeny for dabbling ducks (*Anas*). *Syst. Biol.* **43**, 369–386 (1994).
- Omland, K. Correlated rates of molecular and morphological evolution. *Evolution* **51**, 1381–1393 (1997).
- Strauss, E. Can mitochondrial clocks keep time? *Science* **283**, 1435–1438 (1999).
- Rambaut, A. & Bromham, L. Estimating divergence dates from molecular sequences. *Mol. Biol. Evol.* **15**, 442–448 (1998).
- Bromham, L., Rambaut, A., Fortey, R., Cooper, A. & Penny, D. Testing the Cambrian explosion hypothesis by using a molecular dating technique. *Proc. Natl Acad. Sci. USA* **95**, 12386–12389 (1998).
- Fortey, R. *Life: An Unauthorised Biography* (HarperCollins, London, 1997).
- Valentine, J. W., Erwin, D. H. & Jablonski, D. Developmental evolution of metazoan bodyplans: The fossil evidence. *Dev. Biol.* **173**, 373–381 (1996).
- Bowring, S. A. *et al.* Calibrating rates of early Cambrian evolution. *Science* **261**, 1293–1298 (1993).
- Wray, G., Levinton, J. S. & Shapiro, L. H. Molecular evidence for deep precambrian divergences among the metazoan phyla. *Science* **274**, 568–573 (1996).
- Ayala, F. J., Rzhetsky, A. & Ayala, F. J. Origin of the metazoan phyla: molecular clocks confirm paleontological estimates. *Proc. Natl Acad. Sci. USA* **95**, 606–611 (1998).
- Kerr, R. A. Earliest animals growing younger? *Science* **284**, 411–412 (1999).
- Raup, D. M. *Mass Extinctions: Bad Genes or Bad Luck?* (Oxford Univ. Press, Oxford, 1991).
- Cooper, A. & Penny, D. Mass survival of birds across the Cretaceous–Tertiary boundary: molecular evidence. *Science* **275**, 1109–1113 (1997).
- Hedges, S. B., Parker, P. H., Sibley, C. G. & Kumar, S. Continental breakup and the ordinal diversification of birds and mammals. *Nature* **381**, 226–229 (1996).
- Vilá, C. *et al.* Multiple and ancient origins of the domestic dog. *Science* **276**, 1687–1689 (1997).
- Krings, M. *et al.* Neandertal DNA sequences and the origin of modern humans. *Cell* **90**, 19–30 (1997).
- Darwin, C. *The Origin of Species by Means of Natural Selection*, 6th edn (with corrections and additions to 1872) (John Murray, London, 1888).
- Bowler, P. J. *Charles Darwin: The Man and His Influence* (Blackwell, Oxford, 1991).
- Gould, S. J. & Eldredge, N. Punctuated equilibria comes of age. *Nature* **366**, 223–227 (1993).
- Williams, G. C. *Natural Selection* (Oxford Univ. Press, Oxford, 1992).
- Hecht, M. K. & Hoffman, A. Why no neo-Darwinism: a critique of paleobiological challenges. *Oxford Surv. Evol. Biol.* **3**, 1–47 (1986).
- Fortey, R. M. & Sepkoski, J. J. Jr Absolute measures of the completeness of the fossil record. *Nature* **398**, 415–417 (1999).
- Mooers, A. Ø., Vamori, S. M. & Schluter, D. Using phylogenies to test macroevolutionary hypotheses of trait evolution. *Am. Nat.* **154**, 249–259 (1999).
- Harvey, P. H. & Pagel, M. *The Comparative Method in Evolutionary Biology* (Oxford Univ. Press, Oxford, 1991).
- Pagel, M. Seeking the evolutionary regression coefficient: an analysis of what comparative methods measure. *J. Theor. Biol.* **164**, 191–205 (1993).
- Ridley, M. *The Explanation of Organic Diversity* (Oxford Univ. Press, Oxford, 1981).
- Maddison, W. P. A method for testing the correlated evolutionary of two binary characters: are gains and losses concentrated on certain branches of a phylogenetic tree. *Evolution* **44**, 539–557 (1990).
- Felsenstein, J. Phylogenies and the comparative method. *Am. Nat.* **125**, 1–15 (1985).
- Pagel, M. A method for the analysis of comparative data. *J. Theor. Biol.* **156**, 431–442 (1992).
- Garland, T. Jr, Harvey, P. H. & Ives, A. R. Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Syst. Biol.* **41**, 18–32 (1992).
- Martins, E. P. & Garland, T. H. Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. *Evolution* **45**, 534–557 (1991).
- Purvis, A., Gittleman, J. L. & Luh, H. K. Truth or consequences—effects of phylogenetic accuracy on two comparative methods. *J. Theor. Biol.* **167**, 293–300 (1994).
- Rolland, C., Danchin, E. & de Fraipont, M. The evolution of coloniality in birds in relation to food, habitat, predation, and life-history traits: a comparative analysis. *Am. Nat.* **151**, 514–529 (1998).
- Ricklefs, R. E. & Starck, J. M. Applications of phylogenetically independent contrasts: A mixed progress report. *Oikos* **77**, 167–172 (1996).
- Price, T. Correlated evolution and independent contrasts. *Phil. Trans. R. Soc. Lond. B* **352**, 519–529 (1997).
- Harvey, P. H. & Rambaut, A. Phylogenetic extinction rates and comparative methodology. *Proc. R. Soc. Lond. B* **265**, 1691–1696 (1998).
- Durham, W. *Coevolution: Genes, Culture, and Human Diversity* (Stanford Univ. Press, Stanford, 1991).
- Holden, C. & Mace, R. A phylogenetic analysis of the evolution of lactose digestion. *Hum. Biol.* **69**, 605–628 (1997).
- Jerison, H. J. *Evolution of the Brain and Intelligence* (Academic, New York, 1973).
- Martin, R. D. Relative brain size and metabolic rate in terrestrial vertebrates. *Nature* **293**, 57–60 (1981).
- Pagel, M. & Harvey, P. H. Taxonomic differences in the scaling of brain on body weight in mammals. *Science* **244**, 1589–1593 (1989).
- Arnason, U., Gullberg, A. & Janke, A. Molecular timing of primate divergences as estimated by two nonprimate calibration points. *J. Mol. Evol.* **47**, 718–727 (1998).
- Purvis, A. A composite estimate of primate phylogeny. *Phil. Trans. R. Soc. Lond. B* **348**, 405–421 (1995).
- Healy, S. D. & Krebs, J. R. Food storing and the hippocampus in Paridae. *Brain Behav. Evol.* **47**, 195–199 (1996).
- Wilson, I. & Balding, D. Genealogical inference from microsatellite data. *Genetics* **150**, 499–510 (1998).
- Larget, B. & Simon, D. L. Markov chain monte carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* **16**, 750–759 (1999).
- Ursing, B. M. & Arnason, U. Analyses of mitochondrial genomes strongly support a hippopotamus–whale clade. *Proc. R. Soc. Lond. B* **265**, 2252–2255 (1998).

Acknowledgements

I thank D. Cox, N. Galtier, G. Laden and A. Rambaut for discussion and access to data. The Leverhulme Trust provided financial support.

The role of the Earth's mantle in controlling the frequency of geomagnetic reversals

Gary A. Glatzmaier^{*†}, Robert S. Coe^{*}, Lionel Hongre^{*} & Paul H. Roberts[‡]

^{*} Earth Sciences Department, University of California, Santa Cruz, California 95064, USA

[†] Institute of Geophysics and Planetary Physics, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

[‡] Institute of Geophysics and Planetary Physics, University of California, Los Angeles, California 90095, USA

A series of computer simulations of the Earth's dynamo illustrates how the thermal structure of the lowermost mantle might affect convection and magnetic-field generation in the fluid core. Eight different patterns of heat flux from the core to the mantle are imposed over the core–mantle boundary. Spontaneous magnetic dipole reversals and excursions occur in seven of these cases, although sometimes the field only reverses in the outer part of the core, and then quickly reverses back. The results suggest correlations among the frequency of reversals, the duration over which the reversals occur, the magnetic-field intensity and the secular variation. The case with uniform heat flux at the core–mantle boundary appears most 'Earth-like'. This result suggests that variations in heat flux at the core–mantle boundary of the Earth are smaller than previously thought, possibly because seismic velocity anomalies in the lowermost mantle might have more of a compositional rather than thermal origin, or because of enhanced heat flux in the mantle's zones of ultra-low seismic velocity.

The palaeomagnetic record of the Earth's magnetic field shows that the dipolar part, which is the dominant structure of the geomagnetic field outside the core, has reversed its polarity several hundred times during the past 160 million years (ref. 1). The reversal durations are relatively short (typically 1,000–6,000 years), compared with the constant polarity intervals between reversals. The intensity of the field decreases significantly during reversals. As global records of the palaeomagnetic field are not available, virtual geomagnetic poles (VGPs) have traditionally been computed from local measurements of the field direction using the equations for a purely dipolar field. Geomagnetic excursions (when VGPs deviate more than 45° in latitude from the geographic poles but do not reverse) tend to occur more frequently than full reversals. For example, there is evidence that as many as 14 excursions may have occurred since the last geomagnetic reversal 780,000 years ago².

Unlike the nearly constant periods of the solar magnetic cycle, geomagnetic polarity intervals vary from a few tens of thousands of years to superchrons, which have lasted tens of millions of years (ref. 1). The average duration of geomagnetic polarity intervals during the past 15 million years is about 200,000 years. The duration of a superchron, however, is roughly the timescale required for significant changes in the thermal structure of the Earth's mantle to take place as a result of subduction of tectonic plates and mantle convection. This observation and some noted correlations between plate tectonics, geomagnetic field intensity and reversal frequency have led to speculations that structural changes in the mantle may be influencing convection and magnetic-field generation in the fluid outer core^{3,4} (the geodynamo). In particular, it has been suggested that changes in both the total heat flow^{5–9} and the pattern of heat flux over the core–mantle boundary (CMB)^{3,5,8,10–15} may affect the geodynamo.

Several laboratory and numerical experiments have been conducted to study the effects of non-uniform thermal boundary conditions on non-magnetic convection^{16–19} and on magnetic convection^{20,21}. Here we present three-dimensional numerical simulations of the geodynamo, designed to study the effects of a non-uniform pattern of heat flux over the CMB.

Model description

Our results were obtained using the Glatzmaier–Roberts geodynamo model^{8,22}. This model produced the first dynamically

self-consistent computer simulation of the geodynamo²³: the three-dimensional, time-dependent thermodynamic, velocity and magnetic fields are solved simultaneously, each constantly feeding back to the others. The simulated magnetic field has a dipole-dominated structure and a spatially and time-dependent westward drift of the non-dipolar structure, both similar to the Earth's. A spontaneous magnetic dipole reversal also occurred in the original simulation²⁴. The latest version of this model employs more geophysically realistic heat-flux and magnetic-field boundary conditions, density stratification, inertia, and both thermal and compositional buoyancy sources²⁵.

To simulate hundreds of thousands of years with an average numerical time step (limited by the numerical solution method) of about 15 days, we can afford only a modest spatial resolution in our three-dimensional spherical-core model. That is, we use all spherical harmonics up to degree and order 21, and Chebyshev polynomials in radius up to degree 48 in the fluid outer core and up to 32 in the solid inner core. Consequently, the effects of the small unresolved turbulent eddies, which certainly exist in the Earth's low-viscosity fluid core, need to be represented by "eddy diffusion" in our model, as they are in models of the Earth's atmosphere and oceans and in other models of the geodynamo^{21,26–29}.

We have compared the solutions here with test simulations that have much greater spatial resolutions (one up to spherical harmonic degree 95 and another up to degree 239) and correspondingly less eddy diffusion. Although we see similar axisymmetric structures, the non-axisymmetric parts of the high-resolution solutions are (not unexpectedly) dominated by smaller-scale features. These resolved small-scale eddies provide additional induction of the large-scale magnetic field and so help maintain a more intense field. For example, the magnetic dipole moments for the results reported here tend to be lower than today's geomagnetic dipole moment (7.8×10^{22} A m²) and some are lower than the lowest estimate of its average value (4×10^{22} A m² over the past 160 million years)³⁰. Our high-resolution tests, on the other hand, produce field intensities that are similar to those of the Earth or greater.

In previous dynamo simulations we have, for simplicity, specified a heat flux from the core to the mantle that is uniform over the CMB. However, seismic tomography of the Earth's mantle³¹ and computer simulations of mantle convection³² suggest that the

temperature in the lower mantle can vary by hundreds of degrees Kelvin over lateral distances of roughly 1,000 kilometres. Convection in the fluid core just below the CMB is much more efficient because of the much smaller viscosity; lateral temperature variations there do not exceed 0.001 K. Consequently, large variations in the radial temperature gradient may occur over the CMB, and large

variations may thus also occur in the conduction of heat from the core to the mantle. This would result in slightly cooler (heavier) core fluid, on average, below the cold mantle and slightly warmer (lighter) fluid below the hot mantle, which modifies buoyancy and pressure-gradient forces throughout the fluid core. These forces, together with Coriolis forces (due to the component of the

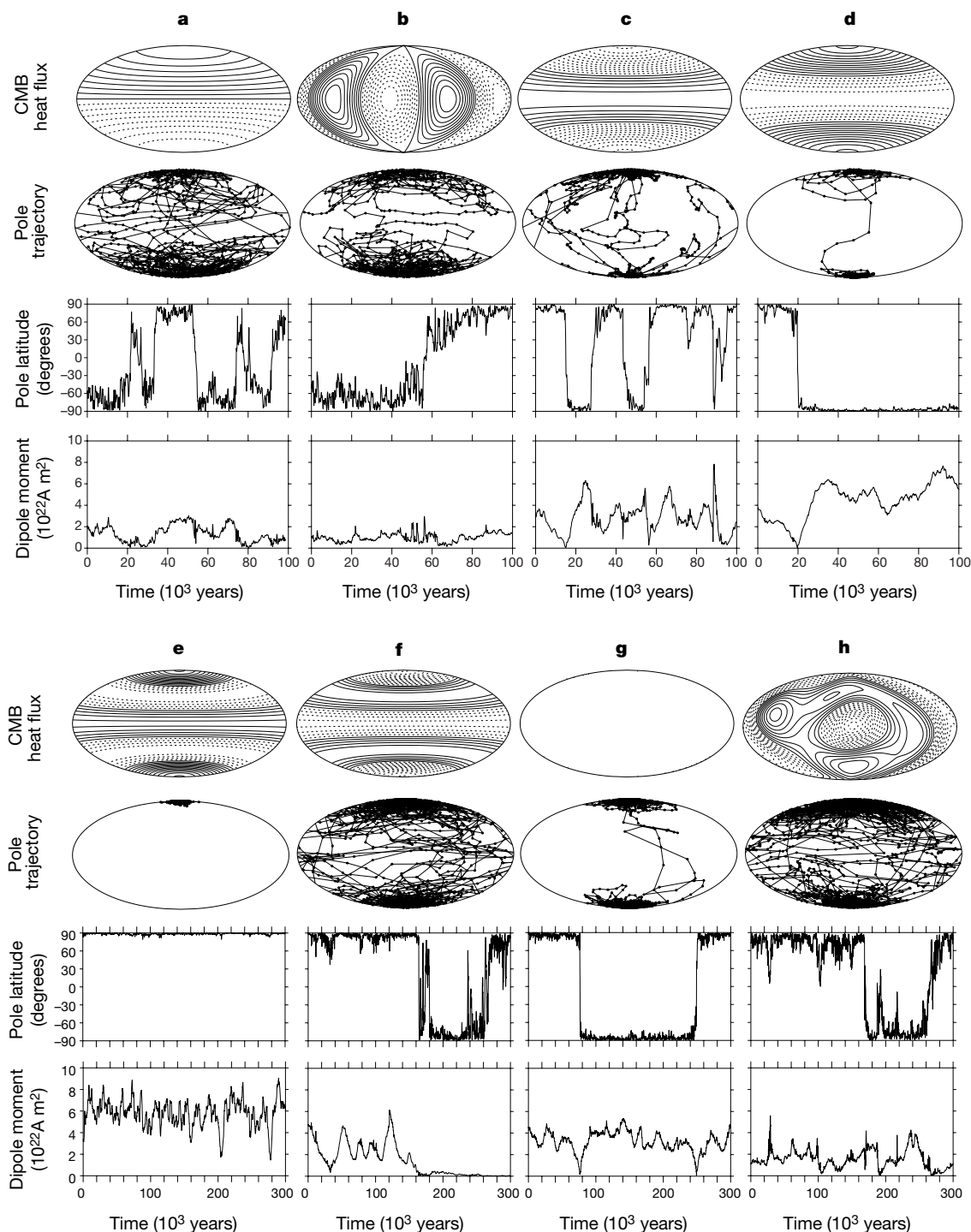


Figure 1 Dynamo simulations. The eight simulations have different imposed patterns of radial heat flux at the core–mantle boundary (CMB). Top row: the pattern of CMB heat flux in Hammer projections over the CMB with the geographical north pole at the top-centre and the south pole at the bottom-centre of each projection. Solid contours represent greater heat flux out of the core relative to the mean; broken contours represent less heat flux. The patterns are proportional to spherical harmonic of degree 1 and order 0 (pattern **a**), of degree 2 and order 2 (pattern **b**), of degree 2 and order 0 (patterns **c** and **d**) and of degree 4 and order 0 (patterns **e** and **f**). The uniform CMB heat-flux case is **g** and the

tomographic case is **h**. Second row: the trajectory of the south magnetic pole of the dipolar part of the magnetic field (observed outside the core) spanning the times indicated in the plots below; the marker dots are about 100 years apart. Third row: plots of the south magnetic pole latitude versus time. Fourth row: plots of the magnitude of the dipole moment versus time. **a–d** span 100,000 years; **e–h** span 300,000 years; the tick marks on the time axes are at intervals of 20,000 years (one dipole magnetic diffusion time). These simulations began long before the 0 times plotted here.

fluid flow perpendicular to the Earth's rotation vector) and Lorentz forces (due to the component of the electric current perpendicular to the magnetic field), drive complicated time-dependent circulations, providing a convective heat flux within the core that accommodates the imposed non-uniform conductive heat flux out of the core at the CMB.

To investigate the possible effects of the mantle's thermal structure on convection and magnetic-field generation in the fluid core, we compare eight dynamo simulations with different, time-independent, patterns of radial heat flux imposed at the CMB of the model. In each case, the total heat flow out of the core is maintained at 7.2×10^{12} W (7.2 TW), only 2.2 TW of which is due to the mean superadiabatic temperature gradient. One of the cases has uniform radial heat flux over the CMB (ref. 22). For the seven non-uniform cases we fix the peak variation of the heat flux (± 0.0446 W m⁻²) at the CMB, relative to the mean, at a value three times greater than the mean superadiabatic heat flux at the CMB; this is based on conservative estimates from mantle convection simulations³². Other than the non-uniform heat-flux boundary condition at the CMB, the model specifications for these simulations are identical to those of the uniform CMB heat-flux case, which was used as an initial condition for the non-uniform cases. In particular, all these cases have the Earth's rotation period and its magnetic dipole diffusion time of 20,000 years.

Results observed at the surface

The eight different CMB heat flux patterns are illustrated in the top row of Fig. 1. Those shown in Fig. 1a–f are simple patterns chosen for this sensitivity study. The uniform CMB heat-flux case (case g) is illustrated in Fig. 1g. The 'tomographic' case (case h), shown in Fig. 1h, is based on the seismic tomography of the lowermost mantle for today's Earth^{8,31} (up to spherical harmonic degree and order four). Cases a–d (Fig. 1a–d) were each run for at least 100,000 years; cases e–h (Fig. 1e–h) were run for at least 300,000 years (15 magnetic dipole diffusion times).

The CMB heat flux in case a is forced to be axisymmetric but highly antisymmetric with respect to the equator; the maximum heat flux out of the core occurs at the geographical north pole and the minimum at the south pole. As observed from outside the core, the direction of the magnetic dipole moment is highly variable, reversing its polarity several times during the 100,000 years displayed in Fig. 1a. The magnitude of the magnetic dipole moment, which includes both the axial and equatorial components, is lower than in several of the other cases and is close to zero during the reversals.

The sensitivity of a longitudinally varying CMB heat flux is tested in case b of Fig. 1; it also varies in latitude but is symmetric with respect to the equator. As in case a, the amplitude of the dipole moment is low, but, unlike case a, only one dipole reversal occurs during the 100,000 years. As seen in Fig. 1b, the pole reversal path seems to follow a longitude band of high CMB heat flux (cold mantle) until reaching the equatorial region where it jumps to the other 'preferred' longitude band, in accord with some palaeomagnetic reversal analyses^{13,33,34} (based on VGPs). However, here, with only one reversal, it could easily be a coincidence. Indeed, some other analyses of palaeomagnetic reversal records suggest that there is no actual longitudinal bias in the distribution of transitional field VGPs (refs 35–37).

The sensitivity of a latitudinally varying CMB heat flux that is axisymmetric and symmetric about the equator is tested in cases c–f. The imposed patterns of CMB heat flux (relative to the mean) are opposite in cases c and d. In both cases, the amplitude of the dipole moment decreases significantly during reversals, but case c reverses more frequently than does case d.

Cases e and f test the effects of greater latitudinal structure in the CMB heat-flux pattern. As seen in Fig. 1, the imposed CMB heat flux for case e peaks at the geographical poles and at the equator and

reaches its minimum at mid-latitude. This is our most efficient case; that is, the amplitude of the dipole moment is, on average, larger than for the other cases. Although the amplitude does have considerable variation, the direction of the dipole moment is always closely aligned with the axis of rotation; no reversal or excursion occurs during the 300,000 years.

Case f, with the opposite pattern of imposed CMB heat flux (relative to the mean), has significantly greater secular variation of the field and a smaller average dipole moment. It reverses after about 160,000 years, but the amplitude of the dipole moment does not recover during the 140,000 years after this first reversal. Before the reversal the magnetic energy integrated throughout the outer core is at least 1,000 times greater than the integrated kinetic energy of the fluid flow (relative to the rotating frame of reference). After this first reversal the magnetic energy is, on average, only about ten times the kinetic energy; it decreases by almost another factor of ten during the second reversal and remains at this level. This is similar to recent results from a number of other dynamo simulations³⁸. Magnetic dipole reversals did not occur in those simulations, but, in order to test the solutions, the magnetic field intensities were manually reduced by a factor of 20 or more during the simulations (within one numerical time step). In a few of the cases the intensities, instead of recovering, remained at roughly the level to which they were artificially reduced. Possibly (as may also apply to case f) the dynamo solutions for those parameters (and, in case f, the CMB conditions) were very close to those for neutral growth of the field. It is also plausible that Venus and/or Mars may have lost strong global magnetic fields in this way.

Our non-uniform CMB heat-flux cases were started from the uniform CMB heat-flux case (Fig. 1g), which was started from our original simulation²³. Case g ran for 500,000 years (25 magnetic dipole diffusion times, 15 million numerical time steps), only 300,000 years of which are displayed. The reversal record for this case is similar to that of the Earth: the axisymmetric part of the field is very stable and dipole-dominated for 230,000 years, then the dipole quickly reverses in ~1,000 years (it takes ~4,000 years for the axial dipole to become dominant again), is stable for the next 170,000 years, then quickly reverses a second time, and is relatively stable for the remaining 100,000 years. This case also shows, more clearly than most of the others, a significant decrease in the dipole moment during reversals (Fig. 1g), as seen in the palaeomagnetic record. A smaller decrease accompanies an excursion that occurs after the time interval displayed in Fig. 1g.

The tomographic case (Fig. 1h), which has an imposed CMB heat flux patterned after today's Earth³¹ (higher heat flux on the 'ring around the Pacific' of high seismic velocity), was chosen to test a possibly more geophysically realistic simulation. We have previously described the first 100,000 years of this case (ref. 8), before the first reversal occurred. The duration of the first reversal is about 6,000 years, whereas the second reversal lasts more than 20,000 years, depending on how the beginning and end of this transition is defined. In addition, several aborted reversals or excursions take place during the simulation. The dipole path of the first reversal strongly overlaps the 'preferred' longitude band of high CMB heat flux that lies underneath the Americas (on the right side of Fig. 1h). Likewise, the density of transitional VGPs from observation points distributed evenly all over the globe also peaks there. The dipole path of the second reversal is much more complex than for the first, sampling many longitudes. So is the density of the transitional VGPs, which has a pronounced low corresponding to the CMB heat-flux low under the Pacific basin and a broad high that extends eastward from 270° E to 135° E, with small peaks centred at 300° E and 35° E. These results and those mentioned above for case b suggest that longitudinally varying CMB heat flux may geographically bias reversal transition paths, but obviously a much longer simulation, with many more reversals than two, would be needed to assess the statistical significance of such correlations. For example,

the two dipole reversal paths of case g, which has no imposed CMB heat-flux bias, just happen to occur in the same longitude band. A more detailed presentation of the characteristics of these reversals will be published elsewhere.

In addition to comparing the frequencies of reversals, we can compare the general structures of the surface fields for the different cases to that of the Earth's. The geomagnetic field at the Earth's surface is traditionally described in terms of gauss coefficients, g_l^m and h_l^m , where l and m are, respectively, the degree and order in the spherical harmonic expansion of the surface field¹. Over the past 5 Myr the Earth's axial quadrupole field, g_2^0 , has had, on average, the same sign as the axial dipole field, g_1^0 , and about 0.04 of its magnitude^{39–43}.

Other than during reversals and excursions, the axial dipole also strongly dominates in our simulations, but the sign relationship between the three main modes is case-dependent. For case a, the axial quadrupole and the axial octupole, g_3^0 , usually have the same sign and reverse together; this reflects the greater field intensity in the northern hemisphere where the heat flux is greater. The dipole and octupole reverse roughly together in cases c and d, maintaining opposite signs in case c and same signs in case d; this reflects the greater field intensity (relative to a pure dipole) maintained in the equatorial region for case c and in the polar regions for case d. In case b, the axial dipole and quadrupole usually have the same sign, opposite the axial octupole, and they all reverse at about the same time. The three modes also reverse approximately together in case g, but the axial quadrupole and octupole nearly always have the same sign, opposite to the sign of the axial dipole. No clear sign relationships persist in cases e, f, h or in our high-resolution versions of case g.

The surface field in our most stable case, e, is strongly dominated by the axial dipole, more so than the Earth's, with the magnitudes of g_2^0 and g_3^0 being on average 0.01 and 0.02 of the magnitude of g_1^0 , respectively. Case f, which has the opposite CMB heat-flux pattern, is our least stable case in terms of maintaining a strong magnetic field. For this case, f, the magnitudes of g_2^0 and g_3^0 are on average 0.05 and 0.04 of g_1^0 before the first dipole reversal. After this reversal, g_1^0 is on average 20 times smaller than it was before; the relative decreases in g_2^0 and g_3^0 are less.

For the other cases, especially c, d and g, which have strong dipolar dominance when not reversing, we see a correlation between the dipolar (W_D) and non-dipolar (W_{ND}) magnetic energy densities (averaged over the surface) during reversals that agrees quite well with the early heuristic model of Cox⁴⁴. Our reversals usually occur when a significant decrease in W_D coincides with an increase in W_{ND} .

Finally, we compare the variability of the surface field to that of the Earth's. Whereas the pole latitude (Fig. 1) provides a measure of the variability of the magnetic dipole, the dispersion of the VGP provides a measure of the directional variability of the entire field at the surface. First the angular standard deviation of the VGPs (from the geographical pole, which is essentially the same as the average magnetic pole) is calculated at over 2,500 sites distributed over the globe, roughly every 100 years, following the usual palaeomagnetic practice of cutting out VGPs that are further than 45° from the geographical poles to exclude reversals and excursions¹. Averaging these in longitude and combining the northern and southern hemispheres provides a VGP dispersion as a function of latitude. An estimate of the Earth's VGP dispersion (ref. 1) over the last 5 Myr (obviously based on much less spatial and temporal resolution than we enjoy) is about 13° at the equator, increasing to about 20° at 65 degrees latitude.

For case a the VGP dispersion is 26° at the equator and decreases slightly with latitude, whereas case b increases from 22° at the equator to 26° poleward. The VGP dispersion for case c is more like the Earth's: 16° at the equator, increasing to 22° at high latitude. Case e is directionally very stable, as the small deviations of its magnetic pole from the geographical pole (Fig. 1e) also suggest; its VGP

dispersion is only 3° at the equator, increasing to about 8° poleward. Very little latitudinal variation is seen in the VGP dispersions for cases d, f, g and h; case d decreases slightly from 10°, case f increases slightly from 14°, case g increases slightly from 11°, and case h increases slightly from 22°. Our higher-resolution version of case g, which has many more degrees of freedom (up to spherical harmonic degree 95), has a VGP dispersion that increases with latitude (8° at the equator, 25° at high latitude), more like the Earth's. However, this is mainly a spatial average because the simulation at this resolution is relatively short.

Dynamics inside the core

The intensity and directional measurements of the field observed above the core are subtle manifestations of the much more intense and complicated field inside the core^{23,24}. We consider, for example, the first reversal of case h. 'Snapshots' of the radial component of the field at the CMB and at what would be the surface of the Earth are displayed in Fig. 2 at roughly 3,000-year intervals, spanning the reversal. The plots for a given snapshot illustrate how the larger-scale structures, such as the dipole, decrease less rapidly with radius, resulting in a much smoother and more large-scale-dominated field at the surface. Corresponding snapshots of the longitudinally averaged field through the interior are also shown. The reversal of the dipolar part of the field begins near the CMB and progresses inward until finally, about 3,000 years after the poloidal field has reversed at the surface (Fig. 2c), the poloidal and toroidal parts of the field reverse (Fig. 2d) inside the 'tangent cylinder', an imaginary cylinder tangent to the solid inner-core equator. Opposite magnetic polarities can exist inside and outside the tangent cylinder²², as seen in Fig. 2c, because the dynamics in these two regions differ significantly. For example, Coriolis forces are always perpendicular to the axis of rotation, whereas buoyancy forces are mostly parallel to the axis inside the tangent cylinder and perpendicular to the axis outside the cylinder; this basic difference in the balance of forces is one of the factors that makes this problem so interesting.

The sequence in Fig. 2 is the opposite of what occurred in our first simulated reversal²⁴. In that reversal the toroidal field reversed first, then the poloidal field inside the tangent cylinder reversed, and finally the poloidal field outside the tangent cylinder reversed. The reversal recently seen in a two-and-a-half-dimensional simulation⁹ (one for which there is insufficient resolution in the third dimension) began at the inner-core boundary. However, for most of the reversals in Fig. 1, the field first reverses outside the tangent cylinder and later inside it. Even the two reversals in case g, with the uniform CMB heat flux, begin outside the tangent cylinder. They both last about 4,000 years when observed at the surface. This includes the time needed to recover dipole dominance; the duration of a reversal at the surface is based on qualitative judgements of its beginning and end times. In addition, a full reversal is not complete until the field inside the tangent cylinder also reverses. For the first reversal of case g, this takes roughly another 2,000 years; the second reversal takes longer, another 9,000 years. The three-dimensional time-dependent details of each reversal presented here differ significantly, as they are likely to do for every geomagnetic reversal that has occurred.

The structure and evolution of the magnetic field is determined by how and where it is twisted and sheared by the fluid flow, which itself is influenced by magnetic (Lorentz) forces. The flow is also greatly influenced by the planetary rotation, the geometry of the inner and outer cores, and the imposed pattern of CMB heat flux. In addition, the flow is driven by thermal and compositional buoyancy sources, which in turn are advected by the flow. This complicated nonlinear system of feedbacks provides an abundant variety of possible solutions that do not lend themselves to simple linear explanations. In general, magnetohydrodynamic instabilities are always occurring, but they usually die away. Once in many attempts, though, an instability continues to grow while the original field polarity decays and the new polarity takes over.

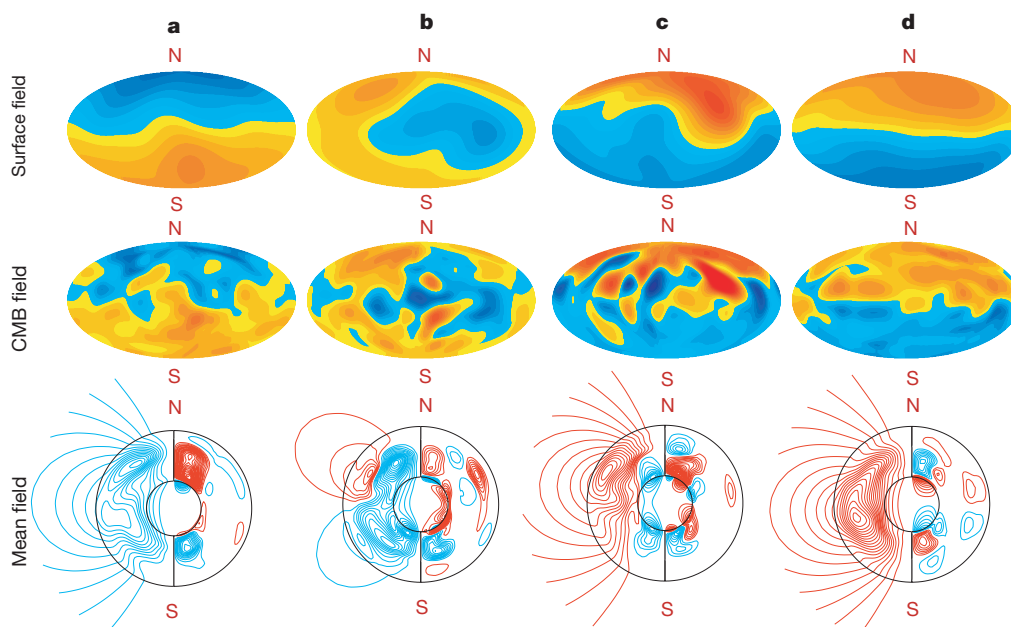


Figure 2 Progression of the magnetic field. A sequence of snapshots is shown of the longitudinally averaged magnetic field through the interior of the core (bottom row), of the radial component of the field at the core–mantle boundary, CMB (middle row), and at what would be the surface of the Earth (top row) displayed at roughly 3,000-year intervals, spanning the first dipole reversal of case h in Fig. 1. In the bottom row of plots, the small circle represents the inner core boundary and the large circle represents the CMB. The poloidal field is shown as magnetic field lines on the left-hand sides of these plots (blue lines are clockwise and red lines are anticlockwise). The toroidal field direction and

intensity are represented as contours (not magnetic field lines) on the right-hand sides (red lines are eastward and blue lines are westward). In the middle and top rows, Hammer projections of the entire CMB and the surface are used to display the radial field (the two different surfaces are displayed as the same size). Yellow shades represent the outward-directed field and blue shades represent the inward-directed field; the surface field, which is typically an order of magnitude weaker, was multiplied by 10 to enhance the colour contrast.

The highly unstable and inefficient magnetic-field generation of case a occurs because the equatorially antisymmetric thermal condition imposed at the CMB is not preferred by the natural dynamics of a rapidly rotating convecting fluid, which attempts to maintain a high degree of thermal symmetry with respect to the equator. Instead, by forcing greater heat flux out of the northern hemisphere, that hemisphere tends to be cooler than the southern hemisphere. This drives hemispheric oscillations in the flow and field amplitudes with a period of ~1,000 years; these oscillations usually lead to a magnetic reversal. However, the reversals seen from outside the core between elapsed model times 20,000 years and 30,000 years (Fig. 1a, third row) actually only occur outside the tangent cylinder. That is, the original polarity of the poloidal and toroidal parts of the field inside the tangent cylinder does not reverse until about time 35,000 years in Fig. 1a when a full magnetic reversal occurs throughout the core. It is likely that cryptochrons (a pair of reversals usually less than 10,000 years apart) and some excursions seen in the palaeomagnetic record² have also occurred only outside the tangent cylinder (S. P. Lund, personal communication).

The relatively large secular variations of the field in cases b and h are due to the longitudinal variations of their CMB heat fluxes²⁰. Convection outside the tangent cylinder, which is mainly in the form of high-wavenumber columnar cells, is continually perturbed by the low-wavenumber thermal boundary condition as the convective pattern propagates westward relative to the mantle⁸. The resulting disturbances in the fluid flow, especially near the CMB, generate disturbances in the magnetic field.

The greater magnetic stability of case d relative to case c and of case e relative to case f indicates a preference for outward convective heat flux in the polar regions, provided by the warm outward-directed part of the thermal wind there⁸. Notice also that the duration of the reversal in case d (Fig. 1) is shorter than those of case c and certainly case f. That is, forcing greater heat flux through the CMB in the polar regions appears to be more compatible with the rotating magnetic convection below the CMB and reinforces the

shear flow on the tangent cylinder, which generates a dipolar field closely aligned with the axis of rotation.

Discussion

The number of reversals that occur in our simulations is far too few for statistical analysis and, in addition, the model should be run with lower viscosity and higher resolution (as computer technology improves), but we can nevertheless make a few observations here. The only simulation that has not yet produced a dipole reversal is case e. Its VGP dispersion is the lowest, and its average dipole moment and dipole dominance are the greatest. Cases d and g also have relatively small VGP dispersions and large average dipole moments; they also have relatively low reversal frequencies and the shortest reversal durations (in terms of pole latitude). The limitations of the short simulation times are apparent in case b, which has a relatively large VGP dispersion but has reversed only once in 100,000 years. However, cases a and c have the highest reversal frequencies and relatively large VGP dispersions.

Correlations like these have been found in the palaeomagnetic record^{1,14,30,44–47}, although there too, exceptions exist⁴⁷. The correlation between high dipole reversal frequency and low dipole moment has also been seen in two-dimensional kinematic dynamo model calculations⁴⁸. Our three-dimensional simulations suggest that the geodynamo is more stable (it has small reversal frequency, secular variation and reversal durations) and more efficient (it has a large dipole moment) when the lateral pattern of diffusive heat flux from the core to the mantle matches the natural time-averaged pattern of convective heat flux deep within the fluid core. This occurs (in our simulations) when the CMB heat flux is axisymmetric and equatorially symmetric, with maxima in the polar regions. These results suggest that superchrons of constant dipole polarity may have occurred under similar conditions and that the pattern of CMB heat flux needs to change significantly to produce a measurable change in reversal frequency.

But how large are the lateral variations in the Earth's CMB heat

flux? Comparing the uniform CMB heat flux case, g, and the tomographic case, h, may be helpful. Although both (so far) have reversal frequencies similar to the Earth's, the excursion frequency may be too high for h and too low for g. However, the VGP dispersion for g, especially the high-resolution version, is more like the Earth's than that for h, and the duration of the second reversal of h is much longer than the more Earth-like ones of g. These preliminary results suggest that the large lateral variations we imposed on the CMB heat flux for case h (inferred from simulations of mantle convection and from large-scale seismic tomography of the lowermost mantle) may be much larger than the Earth's core in fact experiences. If this were the case, our results support two plausible hypotheses (or a combination of them) that have been discussed previously. The seismic velocity anomalies measured in the lowermost mantle may be more a compositional effect than a thermal effect. Or, the CMB heat flux below warm mantle (slow seismic velocity, where we assumed minimum heat flux) may be enhanced by greater mantle thermal conductivity (due to iron enrichment) and possibly by small-scale convection of partial melt corresponding to the ultra-low seismic velocity zones^{49,50} measured in these regions. □

Received 16 June; accepted 10 September 1999.

1. Merrill, R. T., McElhinny, M. W. & McFadden, P. L. *The Magnetic Field of the Earth: Paleomagnetism, the Core, and the Deep Mantle* (Academic, San Diego, 1996).
2. Lund, S. P. *et al.* Geomagnetic field excursions occurred often during the last million years. *Eos* **79**, 178–179 (1998).
3. Cox, A. & Doell, R. R. Long period variations of the geomagnetic field. *Bull. Seismol. Soc. Am.* **54**, 2243–2270 (1964).
4. Vogt, P. R. Changes in geomagnetic reversal frequency at times of tectonic change: evidence for coupling between core and upper mantle processes. *Earth Planet. Sci. Lett.* **25**, 313–321 (1975).
5. Jones, G. M. Thermal interaction of the core and the mantle and long-term behavior of the geomagnetic field. *J. Geophys. Res.* **82**, 1703–1709 (1977).
6. Loper, D. E. & McCartney, K. Mantle plumes and the periodicity of magnetic field reversals. *Geophys. Res. Lett.* **82**, 1703–1709 (1977).
7. McFadden, P. L. & Merrill, R. T. Lower mantle convection and geomagnetism. *J. Geophys. Res.* **89**, 3354–3362 (1984).
8. Glatzmaier, G. A. & Roberts, P. H. Simulating the geodynamo. *Contemp. Phys.* **38**, 269–288 (1997).
9. Sarson, G. R. & Jones, C. A. A convection driven geodynamo reversal model. *Phys. Earth Planet. Inter.* **111**, 3–20 (1999).
10. Hide, R. On the Earth's core-mantle interface. *Q. J. R. Meteorol. Soc.* **96**, 579–590 (1970).
11. Goodacre, A. K. An intriguing empirical correlation between the Earth's magnetic field and plate motions. *Phys. Earth Planet. Inter.* **49**, 3–5 (1987).
12. Bloxham, J. & Gubbins, D. Thermal core-mantle interactions. *Nature* **325**, 511–513 (1987).
13. Laj, C., Mazaud, A., Weeks, R., Fuller, M. & Herrero-Bervera, E. Geomagnetic reversal paths. *Nature* **351**, 447 (1991).
14. McFadden, P. L. & Merrill, R. T. Fundamental transitions in the geodynamo as suggested by paleomagnetic data. *Phys. Earth Planet. Inter.* **91**, 253–260 (1995).
15. Gallet, Y. & Hulot, G. Stationary and nonstationary behavior within the geomagnetic polarity timescale. *Geophys. Res. Lett.* **24**, 1875–1878 (1997).
16. Hart, J. E., Glatzmaier, G. A. & Toomre, J. Space-laboratory and numerical simulations of thermal convection in a rotating hemispherical shell with radial gravity. *J. Fluid Mech.* **173**, 519–544 (1986).
17. Bolton, E. W. & Saylor, B. S. The influence of lateral variations of thermal boundary conditions on core convection: Numerical and laboratory experiments. *Geophys. Astrophys. Fluid Dyn.* **60**, 369–370 (1991).
18. Zhang, K. & Gubbins, D. On convection in the earth's core driven by lateral temperature variations in the lower mantle. *Geophys. J. Int.* **108**, 247–255 (1992).
19. Sun, Z.-P., Schubert, G. & Glatzmaier, G. A. Numerical simulations of thermal convection in a rapidly rotating spherical shell cooled inhomogeneously from above. *Geophys. Astrophys. Fluid Dyn.* **75**, 199–226 (1994).
20. Olson, P. & Glatzmaier, G. A. Magnetoconvection and thermal coupling of the Earth's core and mantle. *Phil. Trans. R. Soc. Lond. A* **354**, 1413–1424 (1996).
21. Sarson, G. R., Jones, C. A. & Longbottom, A. W. The influence of boundary region heterogeneities on the geodynamo. *Phys. Earth Planet. Inter.* **101**, 13–32 (1997).
22. Glatzmaier, G. A. & Roberts, P. H. An anelastic evolutionary geodynamo simulation driven by compositional and thermal convection. *Physica D* **97**, 81–94 (1996).
23. Glatzmaier, G. A. & Roberts, P. H. A three-dimensional convective dynamo solution with rotating and finitely conducting inner core and mantle. *Phys. Earth Planet. Inter.* **91**, 63–75 (1995).
24. Glatzmaier, G. A. & Roberts, P. H. A three-dimensional self-consistent computer simulation of a geomagnetic field reversal. *Nature* **377**, 203–209 (1995).
25. Braginsky, S. I. & Roberts, P. H. Equations governing convection in Earth's core and the geodynamo. *Geophys. Astrophys. Fluid Dyn.* **79**, 1–97 (1995).
26. Kuang, W. & Bloxham, J. An Earth-like numerical dynamo model. *Nature* **389**, 371–374 (1997).
27. Christensen, U., Olson, P. & Glatzmaier, G. A. A dynamo model interpretation of geomagnetic field structures. *Geophys. Res. Lett.* **25**, 1565–1568 (1998).
28. Busse, F. H., Grote, E. & Tilgner, A. On convection driven dynamos in rotating spherical shells. *Studia Geophys. Geodyn.* **42**, 1–6 (1998).
29. Sakuraba, A. & Kono, M. Effect of the inner core on the numerical solution of the magneto-hydrodynamic dynamo. *Phys. Earth Planet. Inter.* **111**, 105–121 (1999).
30. Juez, M. T., Tauxe, L., Gee, J. S. & Pick, T. The intensity of the Earth's magnetic field over the past 160 million years. *Nature* **394**, 878–881 (1998).
31. Su, W.-J., Woodward, R. L. & Dziewonski, A. N. Degree-12 model of shear velocity heterogeneity in the mantle. *J. Geophys. Res.* **99**, 6945–6980 (1994).
32. Tackley, P. J., Stevenson, D. J., Glatzmaier, G. A. & Schubert, G. Effects of multiple phase transitions in a 3-D spherical model of convection in the Earth's mantle. *J. Geophys. Res.* **99**, 15,877–15,901 (1994).
33. Clement, B. M. & Kent, D. V. A southern hemisphere record of the Matuyama-Brunhes polarity reversal. *Geophys. Res. Lett.* **18**, 81–84 (1991).
34. Hoffman, K. A. Dipolar reversal states of the geomagnetic field and core-mantle dynamics. *Nature* **359**, 789–794 (1992).
35. McFadden, P. L., Barton, C. E. & Merrill, R. T. Do virtual geomagnetic poles follow preferred paths during geomagnetic reversals? *Nature* **361**, 342–344 (1993).
36. Prevot, M. & Camps, P. Absence of preferred longitude sectors for poles from volcanic records of geomagnetic reversals. *Nature* **366**, 53–57 (1993).
37. Quidelleur, X. & Valet, J.-P. Paleomagnetic records of excursions and reversals: Possible biases caused by magnetization artefacts. *Phys. Earth Planet. Inter.* **82**, 27–48 (1994).
38. Christensen, U., Olson, P. & Glatzmaier, G. A. Numerical modeling of the geodynamo: A systematic parameter study. *Geophys. J. Int.* **138**, 393–409 (1999).
39. Cox, A. The frequency of geomagnetic reversals and the symmetry of the non-dipole field. *Rev. Geophys. Space Phys.* **13**, 35–51 (1975).
40. Merrill, R. T. & McElhinny, M. W. Anomalies in the time averaged magnetic field and their implications for the lower mantle. *Rev. Geophys. Space Phys.* **15**, 309–323 (1977).
41. Quidelleur, X., Valet, J.-P., Courtillot, V. & Hulot, G. Long-term geometry of the geomagnetic field for the last 5 million years—an updated secular variation database. *Geophys. Res. Lett.* **21**, 1639–1642 (1994).
42. Johnson, C. & Constable, C. The time-averaged field as recorded by lava flows over the past 5 Myr. *Geophys. J. Int.* **122**, 489–519 (1995).
43. McElhinny, M. W., McFadden, P. L. & Merrill, R. T. The time-averaged field 0–5 Ma. *J. Geophys. Res.* **101**, 25007–25027 (1996).
44. Cox, A. Lengths of geomagnetic polarity intervals. *J. Geophys. Res.* **73**, 3249–3260 (1968).
45. Irving, E. & Pullaiah, G. Reversals of the geomagnetic field, magnetostratigraphy, and relative magnitude of secular variation in the Phanerozoic. *Earth Sci. Rev.* **12**, 35–64 (1976).
46. Pal, P. C. & Roberts, P. H. Long-term polarity stability and strength of the geomagnetic dipole. *Nature* **331**, 702–705 (1990).
47. Tauxe, L. & Hartl, P. 11 million years of Oligocene geomagnetic field behavior. *Geophys. J. Int.* **128**, 217–229 (1997).
48. Olson, P. & Hagee, V. L. Geomagnetic polarity reversals, transition field structure, and convection in the outer core. *J. Geophys. Res.* **95**, 4609–4620 (1990).
49. Garnero, E. J. & Helmberger, D. V. Seismic detection of a thin laterally varying boundary layer at the base of the mantle beneath the central-Pacific. *Geophys. Res. Lett.* **23**, 977–980 (1996).
50. Lay, T., Williams, Q. & Garnero, E. J. The core-mantle boundary layer and deep Earth dynamics. *Nature* **392**, 461–468 (1998).

Acknowledgements

We thank R. T. Merrill for suggesting this numerical study. This work was supported by the Institute of Geophysics and Planetary Physics, the Los Alamos LDRD program, the University of California Research Partnership Initiatives program, the NSF Geophysics program and the NASA HPCC/ESS Grand Challenge program. Computing resources were provided by the Los Alamos Advanced Computing Laboratory, the San Diego Supercomputing Center, the Pittsburgh Supercomputing Center, the National Center for Supercomputing Applications, the Texas Advanced Computing Center, the Goddard Space Flight Center, and the Marshall Space Flight Center.

Correspondence and requests for materials should be addressed to G.A.G. (e-mail: glatz@es.ucsc.edu).

Formation of the radio jet in M87 at 100 Schwarzschild radii from the central black hole

William Junor*, John A. Biretta† & Mario Livio‡

* Department of Physics and Astronomy, University of New Mexico, 800 Yale Boulevard, NE, Albuquerque, New Mexico 87131, USA

† Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, Maryland 21218, USA

Massive galaxies often are the source of well collimated jets of material that flow outwards for tens to hundreds of kiloparsecs from the regions surrounding the presumed black holes at their centres. The processes by which the jets are formed and collimated have been important problems for many years¹, and observations have hitherto had insufficient spatial resolution to investigate the length scales associated with these processes^{2,3}. Here we report observations at 43 GHz of the inner regions of the nearby active galaxy M87. The data show a remarkably broad jet having an 'opening angle' of $\sim 60^\circ$ near the centre, with strong collimation of the jet occurring at ~ 30 –100 Schwarzschild radii (r_s) from the black hole: collimation continues out to $\sim 1,000 r_s$. These results are consistent with the hypothesis that jets are formed by an accretion disk around the central black hole, which is threaded by a magnetic field⁴.

The E0 galaxy M87 at the centre of the Virgo cluster contains one of the nearest⁵ (14.7 Mpc from Earth) extragalactic jets^{6–9}. Hubble Space Telescope spectroscopy of its nucleus has given strong evidence for a rapidly rotating ionized gas disk at its centre^{10–12}, from which the presence of a central black hole is inferred, the mass of which is about 3×10^9 solar masses (for which $r_s = 2GM/c^2 \approx 0.0003$ pc, where G is the gravitational constant,

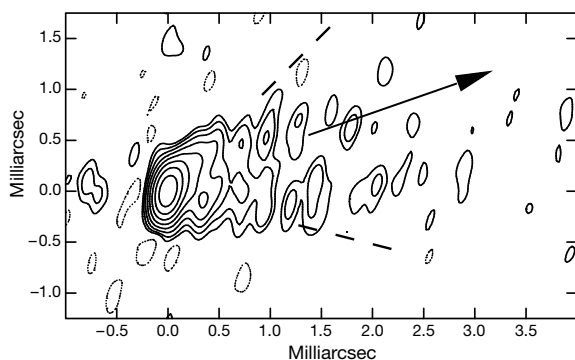


Figure 1 Image of the nucleus of M87 at 43.237 GHz on 3 March 1999. The synthesized beam is $0.33 \text{ mas} \times 0.12 \text{ mas}$ (1 mas is 0.071 pc at the distance of M87) with the major axis in position angle -12.3° . The peak brightness in the image is 228 mJy per beam and the r.m.s. noise in the image well away from the bright structure is 0.38 mJy per beam. Contours are plotted at $-1, 1, 2, 4, 8, 16, 32, 64$ and 128 mJy per beam . The solid arrow indicates the direction of the $20''$ jet, while the dashed lines indicate the position angles of the limb-brightened structure within 1 mas of the core. The antenna array consisted of the 10 telescopes of the Very Long Baseline Array (VLBA), 13 telescopes of the Very Large Array (VLA) phased together, and telescopes located at Effelsberg (Germany), Medicina (Italy), Metsahovi (Finland), Onsala (Sweden) and Yebes (Spain). Left circular polarization data were recorded at each telescope using 8 channels of 8 MHz bandwidth and 2 bit sampling. The data were correlated at the VLBA correlator, and later transferred to the AIPS package for calibration of the complex visibilities and imaging in a standard manner²⁵. The complex visibility data have been weighted by the inverse fourth power of the signal-to-noise ratio of the complex visibility to improve the contribution of the higher spatial frequencies to the image.

M is the black-hole mass, and c the speed of light; the Schwarzschild radius is the radius of the event horizon, the 'boundary' of a black hole. The proximity of M87, together with this large inferred black hole mass (and hence large r_s), make M87 an ideal target for studies of the initial jet-formation process. We note that jets are also associated with young stellar objects and stellar-mass black holes within our own Galaxy^{13,14}, but the small mass and small implied physical scale makes their formation regions extremely difficult to probe.

Centimetre-wavelength very-long-baseline interferometry (VLBI) has been used to study the M87 jet on parsec scales^{15–17}. At these resolutions the jet is quite narrow, and has an opening angle comparable to that measured on kiloparsec scales. To investigate the jet formation and to probe even finer scales, we have recently made VLBI observations with a global array of radio telescopes at a wavelength of 7 mm using the best available receivers and recording equipment. The principal target, M87 (3C274, J1230 + 1223), and a number of calibrator sources were observed with a global array of radio telescopes on 3 March 1999. The resulting image for M87 is shown in Figs 1 and 2. This image has unprecedented resolution^{3,17}, particularly in the direction across the jet's axis, thus allowing study of the jet's collimation.

The core of M87 is clearly detected, as well as jet emission extending out to some 2 milliarcseconds (mas) from the core. The jet appears to be strongly limb-brightened; most of the jet emission of the inner 1 mas extends along ridgelines at position angles of $\sim 315^\circ$ and $\sim 256^\circ$ relative to the core (Fig. 1). These smooth ridgelines of emission extending directly away from the core (especially the northwest limb) underscore interpretation of the jet structure as a cone with a well defined opening angle. We have found very good consistency between this new image and those from the earlier 1.3-cm-wavelength¹⁷ observations with lower resolution, and from an earlier 7-mm global VLBI programme with poorer sensitivity (W.J., J.A.B. and M.L., manuscript in preparation).

The most remarkable feature of the image we report here is the rapid broadening of the jet opening angle as the core is approached on scales below 1 mas (0.1 pc). The jet opening angle, as defined by

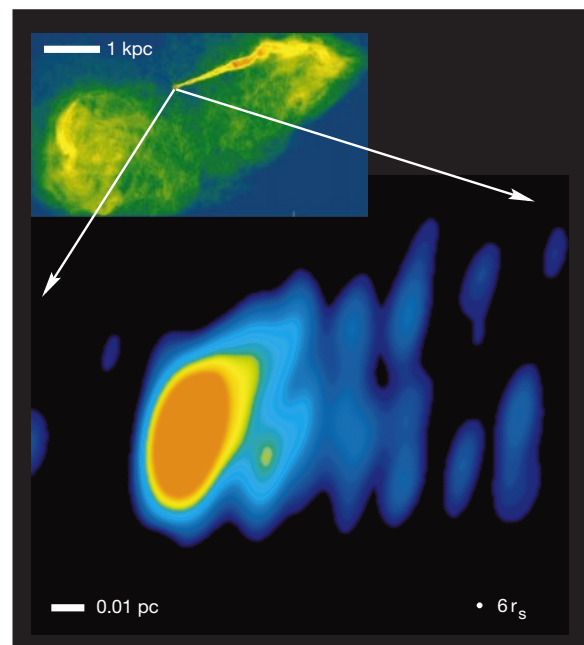


Figure 2 Pseudo-colour rendition of the nucleus of M87 at 43 GHz on 3 March 1999. See Fig. 1 for details. The filled white circle at lower right indicates $6r_s$, which is the diameter of the last stable orbit around a non-rotating black hole. The inset (top left) is a 15-GHz VLA image illustrating the large-scale jet.

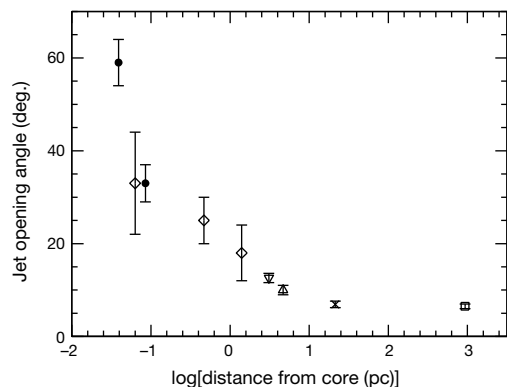


Figure 3 Jet full opening angle as function of distance from the core for M87. Results for a gaussian full-width at quarter-maximum (FWQM) fit¹⁵ are given where jet width is not well-resolved. Uncertainties reflect the effect of image noise on the width measurement. Data are taken from references as follows: filled circles, this paper ($r = 0.1$ pc point using tapered image); open diamonds, ref. 17 FWQM; open inverted triangle, ref. 15 FWQM (new fit assuming zero jet width at core); open triangle, VSOP satellite data plus VLBA data at 5 GHz (W.J., J.A.B., F. Owen and M. Begelman, unpublished results); cross, VLBA 609 MHz FWQM (W.J. and J.A.B., unpublished results); open square, ref. 18.

the limb-brightened structures, is $\sim 60^\circ$ on scales < 0.5 mas (0.04 pc). We believe this to be the broadest angle seen in any extragalactic jet. The angular width of the jet on different size scales has been plotted as a function of radial distance from the core in Fig. 3. It is clear that the jet is progressively wider closer to the core, and that the jet does not adopt its final configuration until a few parsecs from the core. Apparently, there is strong collimation of the jet occurring on scales ≥ 0.04 pc from the central 'engine', which corresponds to $\geq 100 r_s$. The region in which the jet is first formed cannot be much larger than our resolution diameter, or $\sim 30 r_s$. This is within an order of magnitude of the innermost stable orbit radius for a (non-rotating) black hole at $3 r_s$.

We note that the evidence for a rapid increase in the opening angle near the core is independent of possible projection effects. While the inclination angle, θ , between the jet axis and the observer will tend to broaden the apparent opening angle by $\sim (\sin \theta)^{-1}$ (where θ lies between 10° and 40° in M87^{18,9}), it does so similarly all along the jet. Also, the ratios between the jet diameter (or width) where it is being collimated and other relevant length scales (r_s and the size of the accretion disk) are also independent of θ . It is unlikely that the inclination angle θ varies in M87 (that is, that the jet is bent) as the projection of the jet on the sky appears remarkably collinear from 0.02 to 2,000 pc.

The central 'engine' which powers active galactic nuclei (AGN) is thought to be a super-massive black hole with an orbiting accretion disk fed by material raining in from the host galaxy. Magneto-hydrodynamic (MHD) models involving accretion disks are currently favoured for the acceleration and collimation of jets^{4,19}. In these models, MHD 'fluid' from the accretion disk is accelerated along poloidal magnetic field lines threading the disk to speeds which are of order of the keplerian speeds. The outflow in these models could also remove angular momentum from the accretion disk^{20,21}. Collimation of the MHD wind is achieved either through 'hoop stresses' where the field becomes predominantly toroidal, or via poloidal collimation, if the magnetic flux is strongest at the disk's outer radius. Beyond the Alfvén surface, as the gas is no longer attached to the field lines, the field gets wound up by the rotation, thus naturally generating a toroidal field.

These observations are, to our knowledge, the first to show that jets in AGN do not reach their final collimated configuration until a distance many tens of Schwarzschild radii from the 'engine'. On the smallest physical scales yet probed in this

AGN (~ 0.04 pc), the observed outflow already is weakly collimated (opening angle ≈ 1 rad) in the direction of the large-scale jet. This gives the ratio of the collimation distance, r_{coll} , to the fiducial scale $r_{\text{coll}}/r_s \approx 100$. This is consistent with a picture in which the jet is formed (that is, accelerated and collimated) by an accretion disk^{22,23,4}. In particular, our observations seem to favour MHD models in which the jet originates from a disk outflow, rather than emanating already collimated from the vicinity of the black hole.

Evidence for limb-brightening in jets is relatively rare, and has previously been detected only on larger scales^{15,24}. We now consider whether the strong limb-brightening on sub-parsec scales observed here is related to the collimation process itself. Limb-brightening suggests a high synchrotron emissivity, and hence higher magnetic field, in the surface of the jet. This would be consistent with a large-scale magnetic field (that is, from a disk) playing an important role in the collimation process. In order to distinguish further among different specific collimation models, we are currently investigating the magnetic field geometry on these small scales with VLBI polarimetry. □

Received 26 July; accepted 13 September 1999.

- Begelman, M. C., Blandford, R. D. & Rees, M. J. Theory of extragalactic radio sources. *Rev. Mod. Phys.* **56**, 255–351 (1984).
- Zensus, A. J. Parsec-scale jets in extragalactic radio sources. *Annu. Rev. Astron. Astrophys.* **35**, 607–636 (1997).
- Tingay, S. J. *et al.* The subparsec-scale structure and evolution of Centaurus A: the nearest active radio galaxy. *Astron. J.* **115**, 960–974 (1998).
- Livio, M. Astrophysical jets: a phenomenological examination of acceleration and collimation. *Phys. Rep.* **311**, 225–245 (1999).
- Jacoby, G. H., Ciardullo, R. & Ford, H. C. Planetary nebulae as standard candles. V—The distance to the Virgo Cluster. *Astrophys. J.* **356**, 332–349 (1990).
- Curtis, H. D. The planetary nebulae. *Publ. Lick Obs.* **13**, 55–74 (1918).
- Biretta, J. A. in *Astrophysical Jets* (eds Burgarella, D., Livio, M. & O'Dea, C.) 263–304 (Cambridge Univ. Press, 1993).
- Biretta, J. A. in *The Radio Galaxy M87* (eds Röser, H. J. & Meisenheimer, K.) (Springer, Heidelberg, in the press).
- Biretta, J. A., Sparks, W. B. & Macchetto, F. Hubble Space Telescope observations of superluminal motion in the M87 jet. *Astrophys. J.* **520**, 621–626 (1999).
- Ford, H. C. *et al.* Narrowband HST images of M87: Evidence for a disk of ionized gas around a massive black hole. *Astrophys. J.* **435**, L27–L30 (1994).
- Harms, R. J. *et al.* HST FOS spectroscopy of M87: Evidence for a disk of ionized gas around a massive black hole. *Astrophys. J.* **435**, L35–L38 (1994).
- Macchetto, F. D. *et al.* The supermassive black hole of M87 and the kinematics of its associated gaseous disk. *Astrophys. J.* **489**, 579–600 (1997).
- Fender, R. P., Bell-Burnell, S. J. & Waltman, E. B. The radio-jet X-ray binaries. *Vistas Astron.* **41**, 3–13 (1997).
- Ray, T. P. & Mundt, R. in *Astrophysical Jets* (eds Burgarella, D., Livio, M. & O'Dea, C.) 145–175 (Cambridge Univ. Press, 1993).
- Reid, M. J., Biretta, J. A., Junor, W., Muxlow, T. W. B. & Spencer, R. E. Subluminal motion and limb brightening in the nuclear jet of M87. *Astrophys. J.* **336**, 112–120 (1989).
- Biretta, J. A. & Junor, W. The parsec-scale jet in M87. *Proc. Natl Acad. Sci. USA* **92**, 11364–11367 (1995).
- Junor, W. & Biretta, J. A. The radio jet in 3C274 at 0.01 pc resolution. *Astron. J.* **109**, 500–506 (1995).
- Biretta, J. A., Zhou, F. & Owen, F. N. Detection of proper motions in the M87 jet. *Astrophys. J.* **477**, 582–596 (1995).
- Ouyed, R., Pudritz, R. E. & Stone, J. M. Episodic jets from black holes and protostars. *Nature* **385**, 409–414 (1997).
- Blandford, R. D. Accretion disc electrodynamics—a model for double radio sources. *Mon. Not. R. Astron. Soc.* **176**, 465–481 (1976).
- Lovelace, R. V. E. Dynamo model of double radio sources. *Nature* **262**, 649–652 (1976).
- Blandford, R. D. in *Astrophysical Jets* (eds Burgarella, D., Livio, M. & O'Dea, C.) 15–33 (Cambridge Univ. Press, 1993).
- Spruit, H. C. in *Cosmical Magnetism* (ed. Lynden-Bell, D.) 33–44 (Kluwer, Dordrecht, 1994).
- Owen, F. N., Hardee, P. E. & Cornwell, T. J. High-resolution, high dynamic range VLA images of the M87 jet at 2 centimeters. *Astrophys. J.* **340**, 698–707 (1989).
- Zensus, J. A., Diamond, P. J. & Napier, P. J. (eds) *Very Long Baseline Interferometry and the VLBA* (ASP Conf. Ser. 82, Astronomical Soc. of the Pacific, San Francisco, 1995).

Acknowledgements

We thank the staff at the participating observatories for their help, and the software developers at NRAO. The Very Large Array and the Very Long Baseline Array are facilities operated by Associated Universities Inc. under agreement with the National Science Foundation. W.J. thanks the National Science Foundation for support. M.L. was supported by NASA.

Correspondence and requests for materials should be addressed to W.J. (e-mail: bjunor@astro.phys.unm.edu).

Controlled growth of hard-sphere colloidal crystals

Zhengdong Cheng*, William B. Russel† & P. M. Chaikin*

* Department of Physics and † Department of Chemical Engineering, Princeton University, Princeton, New Jersey 08540, USA

Three-dimensional ordered colloidal systems with lattice constants comparable to the wavelength of visible light might find important application as photonic crystals¹, optic filters and switches², and chemical sensors³. Colloidal crystallization has been actively studied^{4–8}, leading to the development of several methods to control the self-assembly of the colloidal particles; examples include colloidal epitaxy⁹ and space-based reduced-gravity techniques^{10,11}. Here we report a method to control the nucleation and growth of hard-sphere colloidal crystals that relies on the use of temperature gradients to define a density gradient. This is somewhat counterintuitive as temperature does not play a role in determining the hard-sphere phase diagram. We obtain hard-sphere single crystals (size ~3 mm) from a sample in a concentration regime that would remain in the liquid state in the absence of a temperature gradient. We expect the method to have applications in controlling the ordering and growth of various 'soft' systems including colloids, copolymers, emulsions and proteins.

Model colloidal hard-sphere suspensions provide a 'laboratory' for the detailed study of many fundamental problems in statistical mechanics, such as crystallization and the glass transition⁵. Hard spheres have no interaction energy, their phase diagram is governed completely by entropy, and hence they are athermal. Counter-intuitively, above a certain volume fraction (0.50) entropy favours the ordered crystalline state over the disordered liquid state. The phase diagram from computer simulations^{12,13} (Fig. 1 top inset) shows a stable colloidal fluid below the freezing volume fraction $\phi_F = 0.494$ and a colloidal crystal above the melting volume fraction $\phi_M = 0.545$. Metastable suspensions with volume fractions in the range $0.494 < \phi < 0.545$ will phase-separate into coexisting crystalline ($\phi = 0.545$) and fluid ($\phi = 0.494$) phases. The phase behaviour has been substantially verified in experiments using sterically stabilized poly(methyl methacrylate) (PMMA) particles suspended in an index-matching solvent^{4,14,15} (that is, in a solvent with a refractive index very close to that of the particles).

Controlled growth of large hard-sphere crystals is of special interest. First of all, the equilibrium structure of the hard-sphere crystal (face-centred cubic, f.c.c.) has only recently been established, and the entropy difference between f.c.c. and hexagonal close-packed (h.c.p.) structures is small^{16–19}. Experiments are affected dramatically by gravity^{10,20} and the dendritic growth instability^{21,22}. The equilibrium f.c.c. structure has yet to be established experimentally. Second, large crystals facilitate the study of lattice dynamics by light scattering. So far, only dilute, charged-sphere crystals have been studied^{23,24}. Last, colloidal crystals have received special attention in the investigation of the structures and properties of photonic crystals, as the particles are comparable in size to the wavelength of visible light and self-assemble into three-dimensional structures^{25,26}. A method for obtaining large, high-quality colloidal crystals is therefore of technological importance.

Large f.c.c. crystals have been grown on a patterned substrate as part of the invention of the colloidal epitaxy technique⁹. This ingenious idea uses gravity to create non-equilibrium (non-thermal) crystals with particles typically in contact with one another. The growth of these crystals is governed by particle sedimentation, and is uncontrolled once the substrate and suspension are prepared. Following the epitaxy idea, the present technique could be

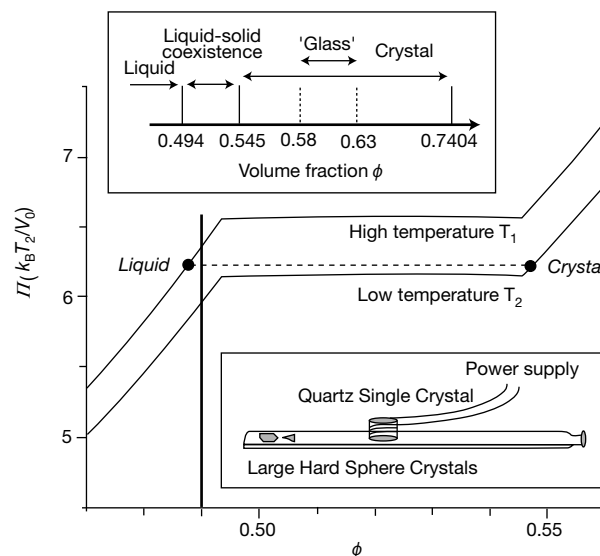


Figure 1 Osmotic pressure as a function of volume fraction. We use the Carnahan–Starling equation and Hall (with $\phi_{\max} = 0.74$) equations of state for $T_2 = 300$ K and $T_1 = 320$ K, with Π in units of $k_B T_2/V_0$, where V_0 is the volume of the sphere, and k_B is the Boltzmann constant. Under each uniform temperature field, the fluid and crystal coexistence regime is $0.494 < \phi < 0.545$. Under a temperature gradient, high-temperature liquid coexists with low-temperature crystal, indicated by the dotted line. The temperature gradient can control the strength of 'supercooling' of metastable fluid, the site of nucleation, and the speed of crystal growth. Top inset, equilibrium phase diagram. Bottom inset, experiment set-up. A chrome–nickel alloy wire heater wound on a quartz crystal cylinder was placed on the upper surface of the flat sample cell. The internal cross-section of the 15-cm-long sample cell was 1 mm $\times$ 1 cm. Initially, the sample was a homogeneous fluid with $\phi \approx 0.49$. The temperature under the heater was 20 °C higher than room temperature (21 °C).

combined with a patterned substrate for nucleation, giving a means to control the growth of oriented equilibrium crystals.

The effect of a temperature gradient on a hard-sphere suspension is shown schematically in Fig. 1, which shows osmotic pressure Π as a function of volume fraction for two different temperatures in the region of the liquid–solid transition. For purely entropic systems, the osmotic pressure is linearly related to the temperature; $\Pi = nk_B T Z(\phi)$, where k_B is Boltzmann's constant, $n = 3\phi/4\pi r^3$ is the particle density for hard spheres of radius r , and $Z(\phi)$ is a known function of volume fraction alone^{27,28}. Now suppose that we have a sample with a volume fraction of 0.493, which at any uniform temperature is a liquid. What happens when a temperature difference is imposed across this sample? Mechanical equilibrium ensures that Π is constant, so $Z(\phi)$ and ϕ vary along the sample, as shown in Fig. 2 for several temperature gradients. The discontinuity indicates the interface between liquid and solid phases. This illustrates how a sample starting in the liquid state in our experiments can be induced to crystallize at the cold end. (Note also that a sample at constant temperature in the coexistence region would have a very different character, with small crystals at $\phi = 0.545$ randomly distributed over the whole sample. The crystallization would take place spontaneously, allowing no control over the position of the nuclei, the growth rate, and the size of the final crystals, which in gravity would sediment and fracture.) By adjusting the strength of the temperature gradient, one may generate a nucleus at the low-temperature end and control growth by changing the gradient to slowly move the liquid–solid interface. If the growth is slow enough, a single crystal may form having the equilibrium structure. However, the thermal-gradient technique only works ideally when gravity generates neither sedimentation nor buoyancy-driven instabilities.

The particles in our experiment are colloidal PMMA spheres,

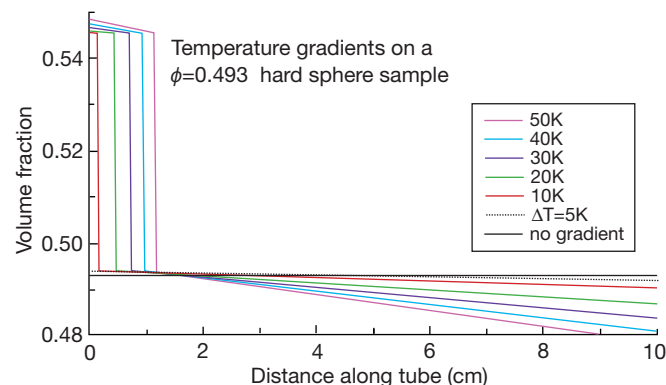


Figure 2 Density profiles. The calculated density (volume fraction ϕ) profile of a sample with average $\phi = 0.493$ is shown for different temperature gradients. The sample would be entirely in the liquid state for any uniform temperature. In a temperature gradient, the denser crystal phase forms at the cooler end of the sample.

with a grafted layer of poly(12-hydroxy stearic acid) about 10 nm thick. Using dynamic light scattering from a dilute sample in decalin, we measured the radius of the spheres to be 349 nm with polydispersity of 5%. The spheres swell slightly in the solvent, which is a mixture of decalin and cycloheptyl bromide with densities of $\rho_{\text{dec}} = 0.896$ and $\rho_{\text{C-B}} = 1.289 \text{ g cm}^{-3}$, respectively. Fortunately, a good density match between the particles and the solvent ($\rho_{\text{PMMA}} = 1.192 \text{ g cm}^{-3}$) also produces reasonable refractive index matching ($n_{\text{PMMA}} = 1.503$, $n_{\text{solvent}} = 1.498$), facilitating static and dynamic light scattering. The quality of density match was checked with a dilute sample, which showed no sign of settling after 12 hours at 1,500g in the centrifuge, indicating an effective gravity smaller than $10^{-3}g$.

Figures 1 (lower inset) and 2 schematically illustrate the set-up and the effects of the temperature gradient. Figure 3a shows a picture of the crystals at the cold end of the sample cell, taken through a close-up lens with white-light illumination. The crystals grew to $\sim 3 \text{ mm}$ in ~ 2 weeks, and remained at this size as long as the temperature gradient was on. The colours result from Bragg scattering. The cool end is fully crystallized, but most of the crystals do not scatter into the camera due to their random orientation relative to the incident light. Figure 3b shows a micrograph of smaller crystals, formed in an experiment with a slightly higher initial volume fraction to demonstrate the crystal morphology. We note that none of the crystals show the dendritic growth instability normally associated with this concentration regime^{10,21,22}. This is an indication of controlled growth.

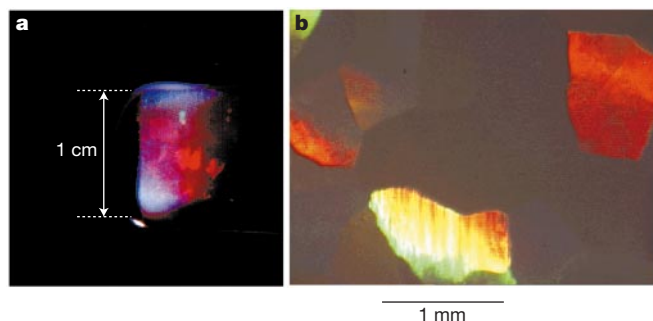


Figure 3 Single colloidal hard-sphere crystals grown by temperature-gradient technique. **a**, Close-up image of the cooler end of the sample cell. The crystals are $\sim 3 \text{ mm}$ in size, and blocky in shape. **b**, Micrograph of crystals from the cooler end of another sample shows the blocky shape of the crystals.

The structure was determined by Bragg scattering from a He–Ne laser beam (wavelength, 633 nm) introduced perpendicular to the cell surface, and diffracted from a single 3-mm crystal onto a white paper screen surrounding the sample. Six symmetrical spots aligned in both forward and backscattering indicate hexagonal order. Rotation of the sample shifted the spots in a manner consistent with Bragg rods from a two-dimensional crystal (or incoherently stacked hexagonal planes). This establishes the structure of the single crystal as random hexagonal close packing (r.h.c.p.), for which the dominant feature in the reciprocal lattice consists of six Bragg rods^{29,30}. Face-centred cubic or h.c.p. structures would have given a very different scattering pattern, as would have a twinned f.c.c. structure. The Bragg scattering angle corresponds to a nearest-neighbour separation in the close-packing plane of 793 nm, which was confirmed by using light at a wavelength of 514.5 nm from an argon laser. All crystals studied show the r.h.c.p. structure. This suggests that our controlled growth was slow enough to suppress the dendritic instability, but still too fast to allow equilibrium growth of the f.c.c. structure. This is consistent with recent calculations of the energy for stacking faults and the rates of growth and annealing to f.c.c.^{17,18}.

Now we consider the kinetics of crystal growth under a temperature gradient. Ideally, the only driving force is the osmotic pressure gradient $\nabla \Pi$ set up by ∇T , which would relax to a constant Π with a volume fraction gradient on a timescale given by gradient diffusion. For our samples, collective or gradient diffusion over the distance of 1 cm would take about 3.5×10^8 seconds or 12 years. Instead, crystallites begin to appear in a sample cell several centimetres in length about two weeks after the temperature gradient is applied.

Although we matched densities to suppress sedimentation, both the colloids and the solvent have significant (and roughly equal) coefficients of thermal expansion. Thus a horizontal temperature gradient can cause convection. The convection velocity U_h can be estimated from $\eta \partial^2 U_h / \partial z^2 \approx \Delta \rho g h / L$, where η is the viscosity of the dispersion, $\Delta \rho$ is the density change of the fluid (colloid + solvent) for a temperature change ΔT , g is the gravitational acceleration, z is the coordinate in the vertical direction, h is the thickness of the sample and L is the distance over which the temperature gradient is applied. Therefore, $U_h \approx \Delta \rho g h^3 / \eta L \approx 2 \times 10^{-5} \text{ m s}^{-1}$, $\Delta \rho = \partial \rho / \partial T \times \Delta T = (0.5 \text{ kg} / (\text{m}^3 \text{ K})) \times 20 \text{ K} \approx 10 \text{ kg m}^{-3}$, $h = 1 \text{ mm}$, $\eta \approx 50(2 \times 10^{-3}) \text{ Pa s}$, and $L \approx 5 \text{ cm}$. The crystal growth rate can be estimated from the diffusion flux as $(\phi_M - \phi_F) dL_{\text{crystal}} / dt = \omega \phi \Delta \mu / l$, where ω is the mobility of the sphere ($D_0 / k_B T$, with D_0 the diffusion constant) and $\Delta \mu$ is the chemical potential difference between the liquid and crystal phases ($\sim k_B T$). The diffusional boundary layer thickness at the end of the cell, $l \approx (h D_0 / U_h)^{1/2}$, is derived from $U_h \partial \phi / \partial z = D_0 \partial^2 \phi / \partial x^2$ where x is the horizontal

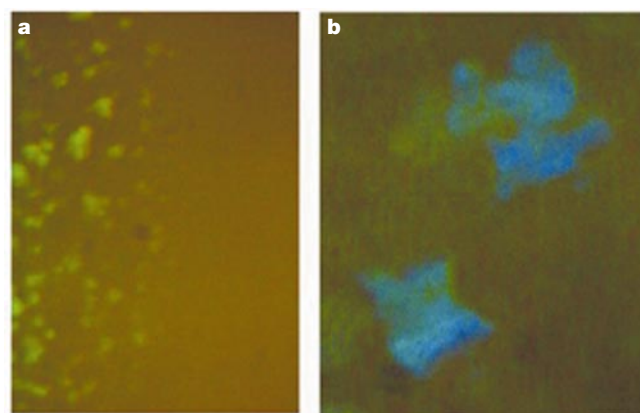


Figure 4 Dendritic crystals. **a**, Moving the heater during growth produces a diffuse liquid–crystal boundary. **b**, A magnified image shows the dendritic growth instability in this region.

coordinate, and the vertical velocity U_v is estimated from continuity as $U_h h \approx U_v L$. Then, $dL_{\text{crystal}}/dt \approx 2 \times 10^{-7} \text{ m s}^{-1}$, which is consistent with the appearance of crystallites observable to the eye only after days. So the buoyancy-driven instability of the fluid in the temperature gradient establishes a convective flow that carries PMMA particles from the hot end to the cool end. The flux of particles deposits some at the cool end due to the low chemical potential of the solid that forms there. A diffusional boundary layer feeds the nucleation and growth process. This crystallization is controlled by the flux of particles along the volume fraction gradient, and can produce very large hard-sphere single crystals if the volume fraction at the cool end is maintained near 0.50 and few nuclei are produced initially.

Moving the heater also has interesting effects. If the velocity is slightly faster than the motion of the growing liquid–solid interface, we observe a more diffuse interface (Fig. 4a), as well as the dendritic growth instability normally found for samples in the coexistence region (Fig. 4b).

We have demonstrated the ability to manipulate the motion of spheres in a suspension using temperature gradients. In the presence of gravity, the convective flow provides the mechanism for macroscopic transport of the particles. The diffusion regime at the cool end of the temperature gradient provides a zone for controlled nucleation and growth of hard-sphere crystals. In the absence of gravity, diffusion fields could be used to control volume fraction more directly in smaller cells. There are other possible uses of temperature gradients. Temperature gradients combined with the colloidal epitaxy technique should allow controlled growth of oriented single crystals. Moving a high-temperature region through a crystalline sample should sweep a low-volume-fraction liquid band through the solid in a technique similar to zone refining for conventional crystals. □

Received 9 April; accepted 9 September 1999.

- Joannopoulos, J. D., Villeneuve, P. R. & Fan, S. Photonic crystals: putting a new twist on light. *Nature* **386**, 143–149 (1997).
- Pan, G., Kesavamoorthy, R. & Asher, S. A. Optically nonlinear Bragg diffracting nanosecond optical switches. *Phys. Rev. Lett.* **78**, 3860–3863 (1997).
- Holtz, J. H. & Asher, S. A. Polymerized colloidal crystal hydrogel films as intelligent chemical sensing materials. *Nature* **389**, 829–832 (1997).
- Pusey, P. N. & van Megen, W. Phase behavior of concentrated suspensions of nearly hard colloidal spheres. *Nature* **320**, 340–342 (1986).
- Pusey, P. N. in *Liquids, Freezing and the Glass Transition* Ch. 10 (eds Hansen, J. P., Levesque, D. & Zinn-Justin, J.) 763–942 (Elsevier, Amsterdam, 1991).
- Poon, W. C. K. & Pusey, P. N. in *Observation, Prediction and Simulation of Phase Transitions in Complex Fluids* (eds Baus, M., Rull, L. F. & Ryckaert, J. P.) 3–51 (Kluwer Academic, Boston, 1995).
- Dinsmore, A. D., Crocker, J. C. & Yodanis, A. G. Self-assembly of colloidal crystals. *Curr. Opin. Colloid Interface Sci.* **3**, 5–11 (1998).
- Grier, D. G. (ed.) From dynamics to device: directed self-assembly of colloidal materials. *MRS Bull.* **23** (10), 21–50 (1998).
- van Blaaderen, A., Ruel, R. & Wiltzius, P. Template-directed colloidal crystallization. *Nature* **385**, 321–324 (1997).
- Zhu, J. *et al.* Crystallization of hard-sphere colloids in microgravity. *Nature* **387**, 883–885 (1997).
- Cheng, S. *Colloidal Hard Sphere Crystallization and Glass Transition*. Thesis, Princeton Univ. (1998).
- Hoover, E. G. & Ree, F. H. Melting transition and communal entropy for hard spheres. *J. Chem. Phys.* **49**, 3609–3617 (1968).
- Woodcock, L. V. Glass transition in the hard sphere model and Kauzmann's paradox. *Ann. NY Acad. Sci.* **371**, 274–298 (1981).
- Phan, S.-E. *et al.* Phase transition, equation of state, and limiting shear viscosities of hard sphere dispersions. *Phys. Rev. E* **54**, 6633–6645 (1996).
- Underwood, S. M., Taylor, J. R. & van Megen, W. Sterically stabilized colloidal particles as model hard spheres. *Langmuir* **10**, 3350–3354 (1994).
- Woodcock, L. V. Entropy difference between the face-centred cubic and hexagonal close-packed crystal structures. *Nature* **385**, 141–143 (1997).
- Bolhuis, P. G. *et al.* Entropy difference between crystal phases. *Nature* **388**, 235–236 (1997).
- Pronk, S. & Frenkel, D. Can stacking faults in hard-sphere crystals anneal out spontaneously? *J. Chem. Phys.* **110**, 4589–4592 (1999).
- Mau, S.-C. & Huse, D. A. Stacking entropy of hard-sphere crystals. *Phys. Rev. E* **59**, 4396–4401 (1999).
- Pusey, P. N. *et al.* Structure of crystals of hard colloidal spheres. *Phys. Rev. Lett.* **63**, 2753–2756 (1989).
- Russel, W. B. *et al.* Dendritic growth of hard sphere crystals. *Langmuir* **13**, 3871–3881 (1997).
- He, Y., Olivier, B. & Ackerson, B. J. Morphology of crystals made of hard spheres. *Langmuir* **13**, 1408–1412 (1997).
- Hurd, A. J. *et al.* Lattice dynamics of colloidal crystals. *Phys. Rev. A* **26**, 2869–2881 (1982).
- Derksen, J. & Van de Water, W. Hydrodynamics of colloidal crystals. *Phys. Rev. A* **45**, 5660–5673 (1992).
- Pradhan, R. D. *et al.* Photonic band structure of bcc colloidal crystals. *Phys. Rev. B* **55**, 9503–9507 (1997).

- Tarhan, I. I. & Watson, G. H. Photonic band structure of fcc colloidal crystals. *Phys. Rev. Lett.* **76**, 315–318 (1996).
- Carnahan, N. F. & Starling, K. E. Equation of state for nonattracting rigid sphere. *J. Chem. Phys.* **51**, 635–636 (1969).
- Hall, K. R. Another hard-sphere equation of state. *J. Chem. Phys.* **57**, 2252–2254 (1972).
- Guinier, A. *X-ray Diffraction* (Freeman, San Francisco, 1963).
- Sanders, J. V. Diffraction of light by opals. *Acta Crystallogr. A* **24**, 427–434 (1968).

Acknowledgements

We thank D. Huse and P. Ségre for discussions. This work was supported by the NASA Microgravity Sciences program.

Correspondence and requests for materials should be addressed to P.M.C. (e-mail: chaikin@puppp.princeton.edu).

High-efficiency multilevel zone plates for keV X-rays

E. Di Fabrizio*, F. Romanato*, M. Gentili†, S. Cabrini†, B. Kaulich‡, J. Susini‡ & R. Barrett‡

* TASC-INFM (National Institute for the Physics of Matter), Elettra Synchrotron Light Source, Lilit Beam-line SS14 km 163.5, Area Science Park, 34012 Basovizza, Trieste, Italy

† Istituto di Elettronica dello Stato Solido, Via Cineto Romano 42, 00156 Rome, Italy

‡ X-Ray Microscopy Beamline, European Synchrotron Radiation Facility, BP220, F-38043 Grenoble Cedex, France

The development of high brilliance X-ray sources coupled with advances in manufacturing technologies has led to significant improvements in submicrometre probes for spectroscopy, diffraction and imaging applications. The generation of a small beam spot size is commonly based on three principles¹: total reflection (as used in optical elements involving mirrors or capillaries), refraction (such as in refractive lenses²) and diffraction. The latter effect is employed in Bragg–Fresnel or Soret lenses, commonly known as Fresnel zone plate lenses. These lenses currently give the best spatial resolution, but are traditionally limited to rather soft X-rays—at high energies, their use is still limited by their efficiency. Here we report the fabrication of high-efficiency, high-contrast gold and nickel multistep (quaternary) Fresnel zone plates using electron beam lithography. We achieve a maximum efficiency of 55% for the nickel plate at 7 keV. In addition to their high efficiency, the lenses offer the advantages of low background signal and effective reduction of unwanted diffraction orders. We anticipate that these lenses should have a significant impact on techniques such as microscopy³, micro-fluorescence⁴ and micro-diffraction⁵, which require medium resolution (500–100 nm) and high flux at fixed energies.

It is only in the last decade, almost a century after the discovery of X-rays (by Röntgen, in 1895), that optical elements with good efficiency and submicrometre resolution have become available. The delays in fabricating such optics were due to the difficulty of realising refraction and reflection at X-ray wavelengths. The physical origin of this technical difficulty is closely linked to the real, phase-shifting part of the refractive index that, for all materials at X-ray wavelengths, is very close to unity (that is, the contrast with respect to the vacuum is very small).

The progress achieved in micro-fabrication has stimulated interest in the fabrication of diffractive optics and, in particular, Fresnel zone plates (FZPs)⁶ (Fig. 1). Many studies considered FZPs to be promising optical elements for microscopy and microprobe applications^{1–18}. This work is related to the development of phase FZP

with 20–40% efficiency in the 5–8 keV energy range^{7–9}, at the Advanced Photon Source at Argonne National Laboratory.

A circular multilevel FZP¹⁰ (Fig. 2) is a good approximation¹¹ of an ideal continuous phase-shifting profile (that is, kinoform) and consists of a series of concentric zone rings, whose radius $r_{n,L}$ is determined by:

$$r_{n,L} = \sqrt{\lambda f \left(2 \frac{l}{L} + n - 2 \right)} \quad (1)$$

where λ is the wavelength, f is the focal length and n is the zone index number (in equation (1), n may have only even values). Each zone is divided into L levels whose index is $l \leq l \leq L$. For a binary FZP ($L = 2$), the zone radius law reduces to the well known binary FZP relation¹⁰, $r_{n,2} = \sqrt{\lambda f n}$. The optical resolution of an FZP is given by $\delta_m = 1.22 \Delta r_{N,L} / m$, where δ_m is the FWHM (full-width at half-maximum) of the Airy disk (the central maximum of the diffraction pattern from a circular lens) at the focal plane for the m th diffraction order, N is the outermost zone index and $\Delta r_{N,L}$ is the outermost zone width. Figure 2 illustrates the working principle of a quaternary FZP ($L = 4$).

The total radiation amplitude delivered by the FZP to the focus can be calculated by using the phasor method¹². The phase shifter of thickness Δt changes the phase by $\phi = 2\pi \Delta t \delta / \lambda$ and attenuates the field amplitude by a factor of $e^{-2\pi \kappa \Delta t / \lambda}$, where δ is the deviation of the real part from unity and κ is the imaginary part of the refractive index, $n_r = 1 - \delta + i\kappa$. When the diffraction condition $\Delta t L \delta = \lambda$ is satisfied, the amplitude A^m is given by:

$$A^m = \frac{C \left[\exp \left(i 2\pi \frac{m}{L} \right) - 1 \right]}{i 2\pi m} \cdot \sum_{l=1}^L \exp \left[- 2\pi \frac{\kappa (l-1)}{\delta L} \right] \cdot \exp \left[+ i 2\pi \frac{(l-1)}{L} (m-1) \right] \quad (2)$$

C is the intensity of the incoming field. The contribution of the diffraction order is taken into account by the index $m = 0, \pm 1, \pm 2, \dots$

From equation (2) the FZP efficiency can be derived, defined as $\eta^m = |A_p^m|^2 / C^2$ for the m th diffraction order. As opposed to the more common binary FZP, the groove asymmetry of the multilevel FZP introduces new phase contributions that interfere in such a way as to concentrate the photon flux in the first diffraction order and to suppress the others almost completely. The imaginary term of the sum in equation (2) accounts for the selection rules of the diffractive order selection. In the limit of zero attenuation ($\kappa = 0$) this term becomes:

$$\sum_{l=1}^L \exp \left[+ i 2\pi \frac{(l-1)}{L} (m-1) \right] = \begin{cases} L & \text{if } (m-1) = 0, \pm L, \pm 2L, \dots \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

The result is that the first order is active and that the other active orders have a periodicity of L . The efficiency is $\eta^m = 2(1 - \cos(2\pi m/L)) / (L/m)^2$.

When absorption in the zones is considered, the theoretical diffraction efficiency of the first order decreases slightly. For a photon energy of 7 keV, it changes from 81.5% to 72% for a FZP made of nickel and to 55% for a gold FZP. Absorption also allows forbidden orders to appear; however, their intensity remains negligible. The zero-order intensity is higher, but $< 1\%$.

On the basis of these results we have fabricated gold (Fig. 1a and b) and nickel (Fig. 1c) FZP lenses for the energy range 5–8 keV. The geometrical characteristics of our FZP at an energy of 8 keV are: focal length, 1 m; diameter, 150 μm ; number of levels $L = 4$; outermost zone-width for the fourth level, 500 nm. A polymethyl methacrylate (PMMA) resist, 1.8 μm thick, was patterned by using electron beam lithography at an energy of 50 keV. The

complete multilevel FZP was manufactured by a three-step process consisting of exposure and electroplating deposition of each level. We developed a custom alignment procedure to reach an accuracy better than 50 nm.

The FZP tests were carried out on the X-ray microscopy beam-line ID21 at ESRF¹³ (European Synchrotron Radiation Facility), which can operate over an energy range of 2–8 keV. The FZP

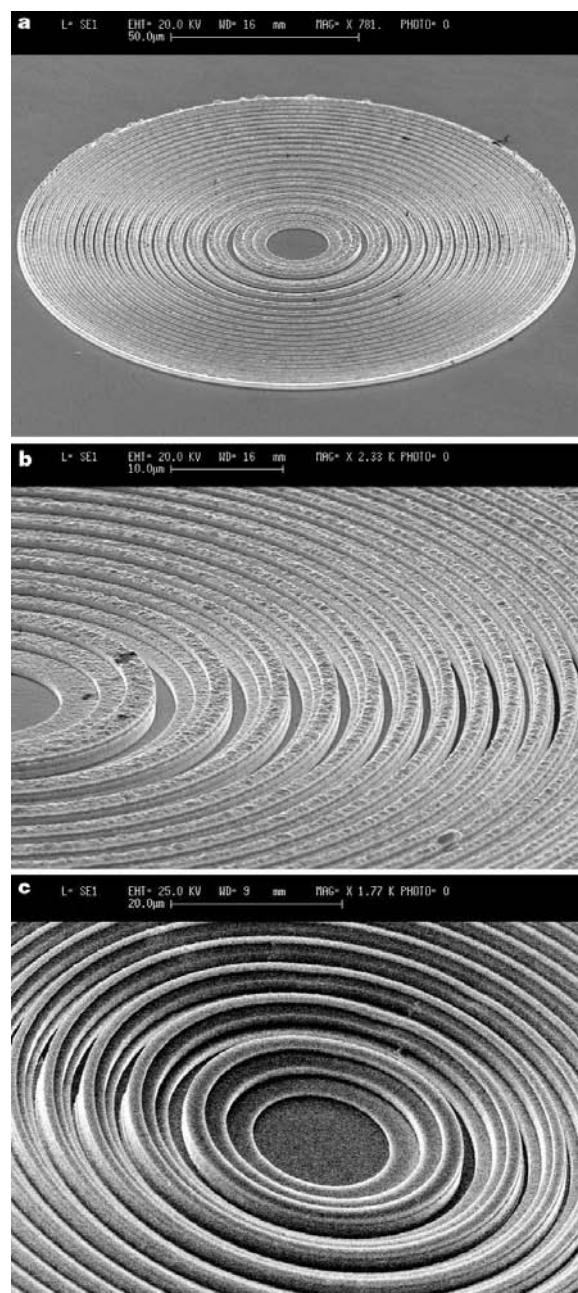


Figure 1 Scanning electron microscope image of quaternary gold (a, b) and nickel (c) Fresnel zone plates. A Fresnel zone plate (FZP) consists of concentric rings designed so that the incident radiation is modulated in amplitude and/or in phase in order to interfere constructively along the optical axis at a point that represents the lens focus. The figure shows FZPs with multilevel profiles. This geometry offers the double advantage of increasing the efficiency (that is, the fraction of the incident intensity delivered to the focus) and of introducing selection rules that highly suppress unwanted diffraction orders. The fabrication of a multilevel FZP requires the deposition and alignment of subsequent levels with an accuracy of a fraction of the desired spatial resolution. The geometrical characteristics designed for 8 keV are: focal length, 1 m; diameter, 150 μm ; number of levels, 4; and outermost zone-width of the fourth level, 500 nm. Scale bars: a, 50 μm ; b, 10 μm ; c, 20 μm .

Table 1 Optical and fabrication design parameters of a quaternary zone plate

Energy (keV)	Theoretical efficiency (%)	Gain	Spot size (nm)	Diameter (mm)	FZP outermost linewidth (nm)	Phase-shifter thickness (μm)
0.76 (low)	40	10^7	100	0.5	50	0.25
7 (medium)	65	10^6	300	0.5	150	1.3
35 (high)	80	10^5	2,000	1	1,000	6.1

efficiency measurements were performed at 5.5, 6.0, 7.0 and 8 keV. The zone plate and the 10 μm pinhole in the focal plane were mutually aligned by means of a stage with a mechanical resolution of 10 nm.

At an energy of 7 keV, the gold FZP provided an efficiency of 38%. The highest efficiency, of 55%, was obtained by the nickel FZP at 7 keV. At 5.5 and 8 keV the efficiency is, however, still higher than 40%. Figure 3a shows the theoretical and experimental efficiency of the nickel FZP. The optical gain (the ratio between the photon density delivered by the FZP at the focus and the photon density impinging on the FZP) is by definition $g = \eta(r_N/(\sigma_x\sigma_y)^{0.5})^2$, where r_N is the FZP diameter and $\sigma_x\sigma_y$ the spot size. The experimental value calculated for the nickel FZP at 7 keV is about 2,500. If we were to work at the diffraction limit by reducing the source size, we would expect a flux gain $g \approx 8,000$. In FZP-based scanning X-ray microscopy, it is common to place an order sorting aperture (OSA) of a few micrometres in the path of the optical axis to decrease the background from spurious orders and thus optimise the image contrast. Figure 3b is a two-dimensional intensity map collected at the focal plane with an OSA with a diameter of 10 μm. The background intensity that passes through an OSA aperture can be written as: $I_{\text{OSA}}^{\pm m_L} = K_L I^{\pm m_L}$, where $I^{\pm m_L}$ are the spurious orders ($m_L \neq 1$). The K_L factor is given by:

$$K_L = \left(\frac{r_{\text{OSA}}}{r_{\text{ZP}}} \right)^2 \sum_{m_L} \frac{1}{(m_L - 1)^2} \quad (4)$$

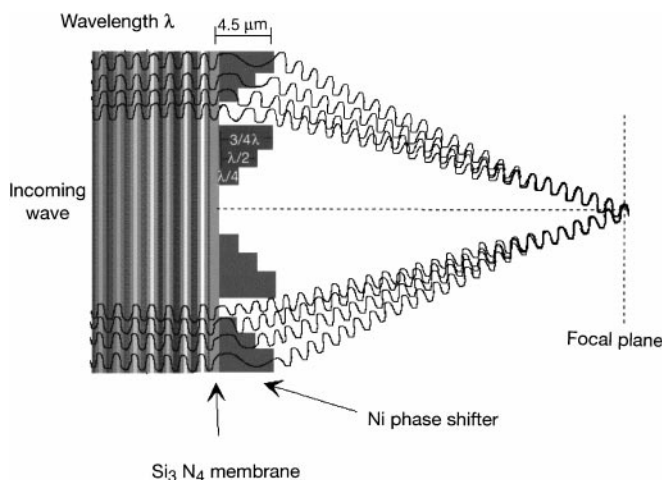


Figure 2 The working principle of a quaternary Fresnel zone plate. The FZP can be based either on amplitude or phase-shift modulation. In the former case, the FZP is necessarily binary. The light impinging on the zones with alternate index is suppressed, by absorption, in such a way that only the wave passing through the open zones will interfere constructively at the focus point. The efficiency of an amplitude FZP cannot be higher than 10%. In the case of phase-shift FZPs, instead of suppressing the part of the incident radiation with the wrong phase, diffraction is performed by phase modulation and allows a binary phase-shift FZP to have an efficiency of up to 40%. In multilevel phase-shift FZPs, diffraction is performed by phase modulation. The FZP described by equation (1) has a periodicity of 2λ in r^2 coordinates. This means that over a period ($\Delta r = 2$) the optical path of the incident radiation increases by λ and the phase by 2π . Within each zone, the level zones represent the interval ($\Delta r = 1$) where the phase of the incident radiation increases by $2\pi/L$. For each level zone, an appropriate thickness of the refractive material (phase-shifter) is then added to the different X-ray optical paths to correct this phase difference. The FZP is fabricated on a Si_3N_4 membrane 2 μm thick.

Equation (4) shows that the quaternary FZP background intensity passing through the OSA is damped by a factor of 10 compared to binary FZP. Reinforcing the diffracted light for a single order makes the use of an OSA unnecessary, and allows multilevel FZPs to be used in a similar way to standard optical lenses. This could be a major advantage in some practical cases in which the conventional geometry (such as the sample–OSA distance) limits the sample environment.

Quaternary FZP can also be used as lenses to condense the beam in an imaging transmission X-ray microscope which requires high flux to be delivered onto the sample with low background. However, within the keV energy range, quaternary FZPs have to be used in conjunction with an additional optical component which allows the numerical aperture of the whole condenser to be matched with the numerical aperture of the X-ray objective lens¹⁴.

To further assess the potential of a quaternary FZP, we discuss the most relevant fabrication parameters: (1) the spatial resolution (the minimum line-width attainable), (2) the aspect ratio (the ratio between the height and the width of a structure), and (3) the alignment between levels. These fabrication figures of merit are

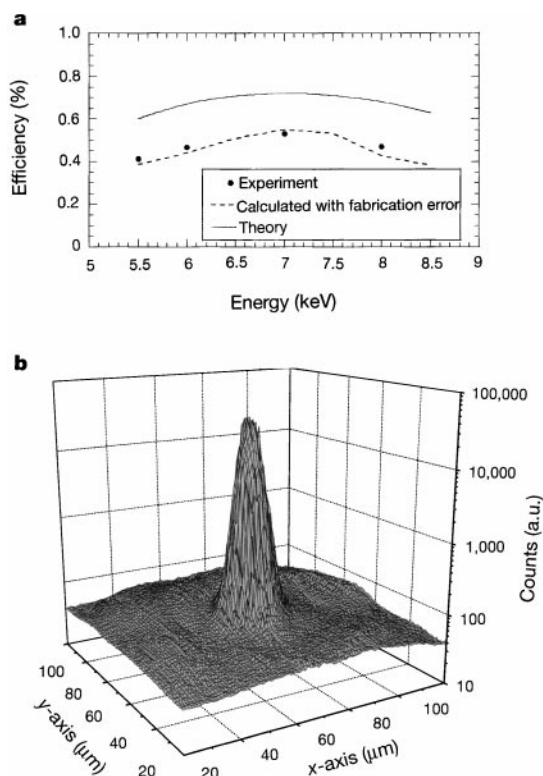


Figure 3 Efficiency and intensity of map of the nickel FZP. **a**, Theoretical and experimental efficiency of a FZP versus energy for the quaternary nickel zone plate. The difference is attributed to fabrication errors, such as electroplating growth, line-width errors and alignment errors. **b**, Two-dimensional intensity map collected at the focal plane through an order sorting aperture with a diameter of 10 μm. The background level is noticeably low compared to the peak intensity (the z axis is a logarithmic scale). This result is due to the multilevel geometry that strongly reduces the intensity in the spurious diffraction orders, and reinforces the first-order peak by a factor of 680 compared to the other orders. a.u., arbitrary units.

related to the optical performances and to the energy at which the optical device is designed to work. In Table 1, the energy region from 500 eV to 50 keV is divided into three parts: low, medium and high energies.

At low energies, the material thickness needed to provide a π -phase shift becomes smaller (200–300 nm), but it is still necessary to achieve high aspect ratios of about 10 if one wishes to fabricate zones at the current lithographic spatial resolution limit^{15,16}. With the latest generation of electron-beam machines and nano-structuring technologies^{17,18}, zone plates with a resolution of 20–50 nm have been fabricated for 200–500 eV X-rays, and alignment accuracy of ~ 10 nm has already been demonstrated⁹. Working at the lithographic limit, the outermost zone width of a quaternary FZP becomes twice as large as that of a binary FZP, thus doubling the optical resolution $\delta_{r,m}$. Therefore, at low energy, quaternary FZPs are complementary to binary FZPs, providing a good trade-off between the reduction of the optical resolution and the increase of the efficiency.

At medium energy, the typical phase-shifter thickness is of the order of one micrometre. Practical aspect ratios of 5–10, that can be achieved by today's very-large-scale integration (VLSI) technology, limit the minimum spatial resolution to 150 nm. At high energy, the phase-shifter thickness further increases to a few micrometres (see Table 1) and, accordingly, the minimum feature size reaches about 1 μ m. In this range, where other optical devices are somewhat limited (although refractive lenses could improve in the future), quaternary FZPs show some unique advantages, especially with respect to their efficiency and optical resolution (see Table 1). Also, multilevel FZPs are compatible at high energy with microdiffraction, where the high focusing length matches the necessary low angular divergence.

With the consolidation of third-generation synchrotrons, multilevel FZPs could be used as focusing elements in research fields where high efficiency and high signal-to-noise ratio are needed at X-ray wavelengths. □

Received 2 June; accepted 9 September 1999.

- Dhez, P., Chevallier, P., Lucatorto, T. B. & Tarrío, C. Instrumental aspects of X-ray microbeams in the range above 1 keV. *Rev. Sci. Instrum.* **70**, 1907–1920 (1999).
- Snigirev, A., Kohn, V., Snigireva, I. & Lengeler, B. A compound refractive lens for focusing high-energy X-rays. *Nature* **384**, 49–51 (1996).
- Kirz, J., Jacobsen, C. & Howells, M. Soft X-ray microscopes and their biological applications. *Q. Rev. Biophys.* **28**, 33–130 (1995).
- Wang, J. et al. X-ray fluorescence correlation spectroscopy: a method for studying particle dynamics in condensed matter. *Phys. Rev. Lett.* **80**, 1110–1113 (1998).
- Rodrigues, W. et al. X-ray microdiffraction studies of an integrated laser-modulator system in *Proc. 6th Conf. on Synchrotron Radiation Instrumentation (SRI97)* (Himeji Citizen Hall, Hyogo, Japan, 1997).
- Tatchyn, R. O. in *X-ray Microscopy* Vol. 43 (eds Smah, G. & Rudolph, D.) 40–50 (Springer Series in Optical Sciences, Springer, Berlin, 1990).
- Lai, B. et al. Hard X-ray phase zone plate fabricated by Lithographic techniques. *Appl. Phys. Lett.* **1**, 1877–1880 (1992).
- Di Fabrizio, E. et al. High-performance multilevel blazed X-ray microscopy Fresnel zone plates fabricating using X-ray lithography. *J. Vac. Sci. Technol. B* **12**, 3979 (1994).
- Yun, W. et al. Nanometer focusing of hard X-rays by phase zone plates. *Rev. Sci. Instrum.* **70**, 2238–2241 (1999).
- Michette, A. G. *Optical Systems for Soft X-rays* (Plenum, New York, 1986).
- Schelokob, I. A. et al. Second generation of grazing-incidence-phase Fresnel zone plates. *Opt. Commun.* **159**, 278–284 (1999).
- Kirz, J. Phase zone plates for X-rays and extreme UV. *J. Opt. Soc. Am.* **64**, 301–309 (1974).
- Barrett, R., Kaulich, B., Oestreich, S. & Susini, J. The scanning microscopy end-station at the ESRF X-ray microscopy beamline. *Proc. SPIE* **3499**, 80–90 (1998).
- Nieman, B. in *X-ray Microscopy and Spectromicroscopy* (eds Thieme, J., Schmahl, G., Rudolph, D. & Umbach, E.) 44–55 (Springer, Chernogolovka, Russia, 1998).
- Spector, S. J., Jacobsen, C. J. & Tennant, D. M. Process optimization for production of sub-20 nm soft X-ray zone plate. *J. Vac. Sci. Technol. B* **15**, 2872–2876 (1997).
- David, C. et al. Low-distortion electron-beam lithography for fabrication of high-resolution germanium and tantalum phase zone plates. *J. Vac. Sci. Technol. B* **13**, 2762–2766 (1995).
- Schneider, G., Schliebe, T. & Aschoff, H. Cross linked polymers for nanofabrication of high resolution zone-plates in nickel and germanium. *J. Vac. Sci. Technol. B* **13**, 2809–2812 (1995).
- Weiss, D., Peuker, M. & Schneider, G. Radiation enhanced network formation in copolymer galvanofoms for diffractive nickel X-ray optics with high aspect ratios. *Appl. Phys. Lett.* **72**, 1805–1807 (1998).

Correspondence and requests for materials should be addressed to E.D.F. (e-mail: difabriz@sci.area.trieste.it).

Influence of environmental changes on degradation of chiral pollutants in soils

David L. Lewis^{*†}, A. Wayne Garrison^{*}, K. Eric Wommack^{*†}, Alton Whittemore^{*}, Paul Steudler[‡] & Jerry Melillo[‡]

^{*} US Environmental Protection Agency, Ecosystems Research Division, National Exposure Research Laboratory, Athens, Georgia 30605, USA

[†] Department of Marine Sciences, University of Georgia, Athens, Georgia 30602, USA

[‡] The Ecosystems Center, Marine Biological Laboratory, Woods Hole, Massachusetts 02542, USA

Numerous anthropogenic chemicals of environmental concern—including some phenoxy acid herbicides, organophosphorus insecticides, polychlorinated biphenyls, phthalates, freon substitutes and some DDT derivatives—are chiral. Their potential biological effects, such as toxicity, mutagenicity, carcinogenicity, and endocrine disrupter activity, are generally enantiomer-selective, and different enantiomers are preferentially degraded (transformed) by micro-organisms in various environments^{1–8}. Here we use field and laboratory experiments to demonstrate that environmental changes in soils can alter these preferences, and to suggest that the preferences shift owing to different groups of related microbial genotypes being activated by different environmental changes. In Brazilian soils, almost all pasture samples preferentially transformed the non-herbicidal enantiomer of dichlorprop ((*RS*)-2-(2,4-dichlorophenoxy)propionic acid), while most forest samples either transformed the herbicidal enantiomer more readily or as rapidly as the non-herbicidal enantiomer. Organic nutrient enrichments shifted enantioselectivity for methyl dichlorprop ((*RS*)-methyl 2-(2,4-dichlorophenoxy)propionic acid) strongly towards preferentially removing the non-herbicidal enantiomer in soils from Brazil and North America, potentially increasing phytotoxicity of its residues relative to that of the racemate. Assessments of the risks chemical pollutants pose to public health and the environment need to take into account the chiral selectivity of microbial transformation processes and their alteration by environmental changes, especially for pesticides as up to 25 per cent are chiral⁹.

Enantiospecific biological effects associated with chiral chemicals are well recognized in the pharmaceutical industry where 50 of the top 100 most widely sold drugs, including barbiturates, ibuprofen, albuterol, and Ritalin, are marketed as single enantiomers to avoid adverse side effects^{8,10}. The high cost of separating enantiomers

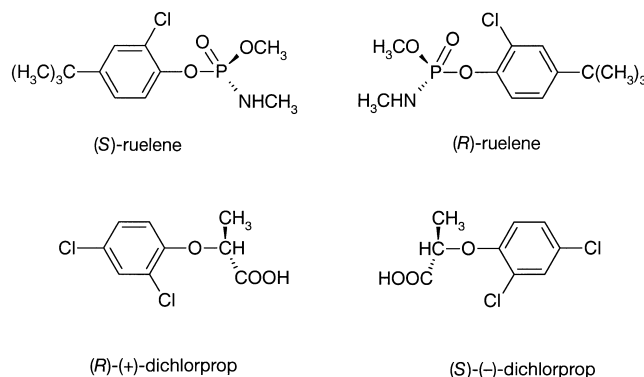


Figure 1 Enantiomers of ruelene and dichlorprop.

Table 1 Effect of deforestation on dichlorprop enantioselectivity in Brazilian soils

	<i>n</i>	<i>t</i> _{1/2} (days)	c.v. (%)	(+) (%)	(-) (%)	(+/-) (%)
Methyl dichlorprop						
Forest	52	0.77	62	7.7	1.9	90
Pasture	75	0.63	49	11	1.7	87
Dichlorprop						
Forest	16	11	36	13	38	50
Pasture	37	16	57	6.7	93	0.0

Soil samples were collected from a forest in Rondonia, Brazil and from six deforested areas (pastures) nearby, ranging 9–47 years in age. Soil samples (1 gram) were amended with herbicides and incubated overnight for methyl dichlorprop and over ~6 months for dichlorprop. Selectivity for different enantiomers ((+) or (-)) was quantified as percentages of soil samples in which differences in the amounts transformed of each enantiomer were ≥15%. Non-selective samples (+/-) had concentration differences between (+) and (-) enantiomers of <15%. Transformation half-lives (*t*_{1/2}) of herbicide racemates were calculated ± a coefficient of variation (c.v.) among a number (*n*) of replicate soil samples. Almost all (93%) of the pasture samples preferentially transformed the non-herbicidal (-) enantiomer, while most forest samples either preferentially transformed the herbicidal enantiomer (13%) or transformed it as rapidly as the (-) enantiomer (50%).

currently prohibits applying this technology to the commercial production of most pesticides and other industrial chemicals to reduce the unwanted effects of racemates as environmental pollutants. Notable exceptions include two herbicides, dichlorprop and mecoprop, in which a limited portion of the total worldwide production includes single enantiomers¹¹.

In the case of chiral pollutants, environmental studies have historically neglected to determine the adverse effects associated with particular enantiomers, and which enantiomers may persist in the environment. Consequently, much of the environmental data that has been collected for chiral pollutants, including many of the chemicals raising the greatest public concern, may have represented only the presence of relatively innocuous enantiomers.

To assess the persistence of different pollutant enantiomers across a wide latitudinal transect, and the possible influence of large-scale environmental changes on that persistence, we investigated enantioselectivity in soils collected from three areas after amending them with ruelene ((*RS*)-4-tert-butyl-2-chlorophenylmethyl-*N*-methyl phosphoramidate), dichlorprop, and methyl dichlorprop (Fig. 1). The sites included an upland plateau in Risdalsheia, Norway about 20 km inland of the North Sea, an 80-year-old mixed deciduous forest (Harvard Forest) in the northeastern US, and an area of the Fazenda Nova Vida about 250 km south of the city of Porto Velho in Rondonia, Brazil. Only the (+) enantiomers of dichlorprop and methyl dichlorprop are herbicidal¹¹ (see Methods). Both enantiomers of ruelene were insecticidal; however, the (+) enantiomer was four times more toxic.

Table 1 shows that tropical deforestation (conversion to pastures) had little or no effect on enantioselectivity or demethylation rates of the racemate of methyl dichlorprop. Demethylation was the first and most rapid step in the transformation process, which quantitatively produced dichlorprop as the reaction product. Soil warming in North America and Norway also had no measurable effect on racemate demethylation rates or enantioselectivity for methyl dichlorprop where field plots were heated 5 °C above ambient temperatures for ~7 years in North America and ~4 years in Norway^{12,13}. Transformation of dichlorprop began after a 6-month lag period and proceeded 20 times more slowly (half-life of ~14 days compared with ~0.7 days for methyl dichlorprop).

Addition of inorganic fertilizers to forest and pasture plots in Brazil had no measurable effect on enantioselectivity for methyl dichlorprop (tested at 2 weeks and at 5 months after fertilization), or to plots in North America fertilized since May 1988. Laboratory enrichments with organic nutrients, which caused soils preferentially transforming the (+) enantiomer to shift to strongly preferentially transforming the inactive (-) enantiomer in samples from North America and Brazil (Table 2), did not have the same effect in samples from Norway where soils already preferentially removed

Table 2 Effect of nutrient enrichment on enantioselectivity for methyl dichlorprop

	<i>n</i>	(+) (%)	(-) (%)	(+/-) (%)	Net preference (%)
Rondonia, Brazil					
Soil	127	9.0	2.0	88	7.0 (+)
Enrichment	340	15	50	35	35 (-)
Harvard Forest, Massachusetts, USA					
Soil	29	45	48	7.0	3.0 (-)
Enrichment	207	3.0	80	17	77 (-)
Risdalsheia, Norway					
Soil	16	0.0	88	13	88 (-)
Enrichment	80	15	74	11	59 (-)

Methods for testing soil were the same as Table 1, except that the soil samples contained beef extract and peptone. Net preferences, which indicated the extent to which soil (or soil + organic nutrients) preferentially transformed an enantiomer, were calculated as the absolute values of differences in percentage of soil samples preferring (+) or (-) enantiomers.

the (-) enantiomer. There, microorganisms preferring the (+) enantiomer may have increased in activity somewhat after nutrient amendments, but most samples (74%) still preferred the (-) enantiomer. Preferential removal of (-) methyl dichlorprop would render residues more phytotoxic than the same concentrations of the racemate, as long as demethylation is a prerequisite step and the (+) dichlorprop enantiomer is not preferentially removed after nutrient enrichment.

Ruelene, like dichlorprop, was transformed by microbial populations only after about a 6-month lag period. The half-life, once transformation began, was ~40 days in Brazilian soils and 20 days in Norwegian soils. Ruelene racemate transformation rates were unaffected by deforestation or soil warming. Enantioselectivity, however, was affected in both cases. In Brazil, deforestation caused soils to shift from preferentially removing the less toxic ruelene (-) enantiomer (22% (+), 67% (-), 11% (+/-), *n* = 17) to exclusively transforming the (+) enantiomer. Soil warming in Norway, on the other hand, caused soils to shift from all samples preferentially removing the ruelene (+) enantiomer to some (22%) preferentially removing the less toxic enantiomer (*n* = 13). Technical obstacles to separating ruelene enantiomers were greater than those faced with dichlorprop; therefore, fewer soil samples were successfully analysed (17 from Brazil and 13 from Norway). Nevertheless, results were consistent with the behaviour of dichlorprop, where environmental changes affected enantioselectivity when long lag periods were required for adaptation.

Delayed microbial transformation of chemical residues is commonly observed and is often assumed to represent the amount of time required for low concentrations of microorganisms responsible for the transformation process to attain significant numbers after using the chemicals as substrates or energy sources. Hollibaugh¹⁴, however, investigated enantioselectivity for D- and L-amino acids metabolized by marine bacteria and found that lag periods could not be explained by growth of sub-populations. Instead, they were consistent with times required for activating metabolically quiescent microbial populations or inducing enzymes for enantiomer-specific uptake or transformation processes.

As environmental changes in this study affected only the two chemicals with long lag periods, sorption, a more rapid process, probably did not play a role. Enantioselective sorption has been ruled out elsewhere as well⁷. Nevertheless, many soils have their genesis with living matter and their organic components may retain chiral centres capable of selectively binding specific enantiomers. Enantioselective sorption to soil organics, if it occurs, could affect the availability of various enantiomers for microbial transformation.

To investigate the diversity and biogeography of bacteria capable of demethylating methyl dichlorprop, small-subunit (16S) ribosomal

RNA (rRNA) genes of 50 soil bacterial isolates were compared using amplified ribosomal DNA restriction analysis (ARDRA)¹⁵ (Fig. 2). ARDRA has proved useful in taxonomic assessments of bacterial culture collections¹⁶ and 16S rDNA clone libraries¹⁷ at genus and species levels.

In general, isolates from Norway and North America formed three large clusters and appeared to be more closely related to one another than to isolates from Brazil. Conversely, Brazilian isolates formed smaller clusters that were more distantly related to one another and to Norwegian and North American clusters. It is plausible that these relationships among bacterial isolates reflect ecological differences between these geographical regions. ARDRA patterns that indicated taxonomic identity between two or more isolates were, in every comparison examined except one (Norway isolates 409-1 versus 416-4), supported by 16S rDNA sequence data demonstrating gene-sequence identity. Furthermore, isolates with identical ARDRA patterns exhibited the same enantioselectivity phenotype regardless of their geographical origins, except in the case of the two isolates noted above.

Although the ARDRA banding pattern was virtually the same for Norway isolates 409-1, which preferentially transformed the (–) enantiomer, and 416-4, which was non-enantioselective, 16S rDNA

sequence data varied by several base pairs. Isolate 409-1 did metabolize the (+) enantiomer, but at a much slower rate than the (–) enantiomer. In fact, with all soil samples and bacterial isolates, both enantiomers for each test chemical were transformed simultaneously, but at different rates. This behaviour is consistent with what has been observed by others where enantioselectivity has been detected in environmental samples.

Our data suggest that microbial enantioselectivity with environmental pollutants is controlled by the activation of metabolically quiescent microbial populations or the induction of enantiomer-specific enzymes, as was the case with amino acids, and that this selectivity follows lines of genetic similarity where different groups of related microbes are activated by different kinds of environmental changes.

With one exception (Norway 409-1), only isolates from Brazil exhibited enantioselectivity with methyl dichlorprop. Among those isolates, 58% were enantioselective and their preferences roughly reflected those observed with soil incubations (Table 2). However, Norwegian and North American isolates did not reflect the (–) enantiomer preference observed in soil incubations from these locations. Most of these isolates were non-enantioselective, but the relatively few genotypes that were enantioselective tended to

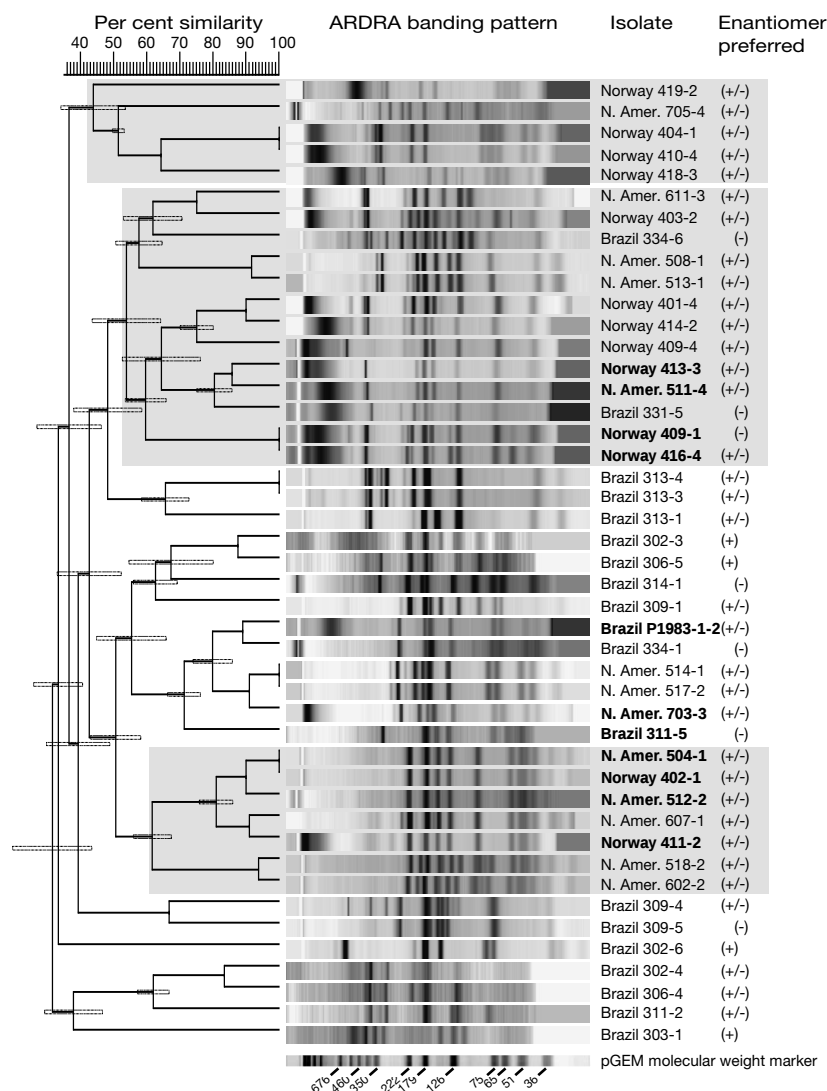


Figure 2 Amplified ribosomal DNA restriction analysis of bacterial isolates. Confidence limits for the placement of nodes are shown with error bars. Molecular weight marker (pGEM) to which band positions were normalized is shown at the bottom. Numbers

indicate size in base pairs of DNA. In general, isolates that enantioselectively transformed methyl dichlorprop from North American and Norwegian soils formed three large clusters (grey boxes) whereas Brazilian isolates formed smaller, more distantly related, clusters.

dominate transformation processes such that soil samples from North America preferred either (+) or (–) enantiomers while almost all of the soil samples from Norway preferred the (–) enantiomer.

These results have important implications for public health and environmental protection. The differing effects of inorganic and organic nutrients on enantioselectivity (Tables 1, 2), for example, raise new questions with respect to land application of processed sewage sludge (biosolids), which transforms pollutants enantioselectively¹⁸. Under a regulation first promulgated by the US Environmental Protection Agency in 1993 (ref. 19), biosolids are used as a commercial fertilizer for agricultural lands heavily treated with pesticides. In the USA, Europe, and elsewhere, governments are promoting the agricultural use of biosolids to replace ocean dumping of municipal wastes. Substituting organic-rich biosolids for inorganic fertilizers may increase microbial transformation rates of some of the active enantiomers of pesticides and thus diminish their effectiveness in the field. On the other hand, persistence may be enhanced for some enantiomers posing a risk to public health and the environment.

This study demonstrates that significant environmental changes, such as tropical deforestation and global warming, may substantially alter the relative persistence of enantiomers of chiral pollutants, exacerbating the adverse effects of some while ameliorating the effects of others. It calls into question the accuracy and future relevance of current risk assessments for numerous pollutants, and underscores the need to better incorporate science in efforts aimed at protecting public health and the environment²⁰. □

Methods

Dichlorprop and ruelene were microbially transformed rapidly enough to process numerous samples, and could also be adequately separated from extracts of soils and microbiological media. Dichlorprop, one of the most widely used herbicides, is applied as the acid and various esters that are readily attacked by common bacterial esterases. Ruelene has been banned in the USA but is still used elsewhere. Chiral Technologies (Exton, Pennsylvania, USA) separated the enantiomers from racemates and methylated the dichlorprop enantiomers.

Brazilian soil samples were collected in March and August 1998 from a forest and an adjacent 11-year-old pasture. Plots in both areas were treated with inorganic fertilizers (~33 kg N per hectare and 13 kg P per hectare). Additional samples were collected from 9-, 15-, 19-, 26- and 47-year-old pastures. Soil-warming samples with corresponding control (unheated) samples were collected in October 1998 from field plots in Harvard Forest, Massachusetts where buried, computer-controlled electrical cables heated soil 5 °C above ambient temperatures, beginning in July 1991. Samples were also collected in September 1998 from field plots in Risdalsheia, Norway that had been similarly heated since May 1994.

Soil samples of 1-gram each were treated with 10-ml aqueous racemic herbicide or insecticide solutions containing ~25 mg l⁻¹ of each enantiomer and incubated at 25 °C overnight for methyl dichlorprop and ~6 months for dichlorprop and ruelene. Half-lives were estimated from changes in pollutant concentrations after transformation began, assuming first-order kinetics. Enantiomer concentrations usually decreased >80%.

Methyl dichlorprop enantiomers were detected by flame ionization (FID) after separation with a 25 m × 0.25 mm wall coated, open tubular fused silica column coated with a 0.25-µm layer of CP ChirasilDex CB (Chrompack, Ravit, New Jersey, USA). Dichlorprop enantiomers were separated by capillary electrophoresis (CE) (15 kV with 25 mM tetraborate buffer and 25 mM trimethyl-β-cyclodextrin) and detected by ultraviolet absorption at 230 nm. Ruelene enantiomers were also separated by CE in the micellar electrokinetic chromatography mode (20 kV with 20 mM tetraborate buffer at pH 8.5 containing 100 mM sodium dodecyl sulphate, 20% acetonitrile, and 40 mM 2-hydroxypropyl-β-cyclodextrin) and detected at 200 nm.

Only samples demonstrating a decrease of ≥15% in racemates (two standard deviations of the experimental error for determining concentrations) were analysed for enantioselectivity. Samples exhibiting enantioselectivity generally had a ≥50% difference between concentrations of the two enantiomers at the final sampling time; non-enantioselective samples (+/–) generally showed no significant differences.

For soils enriched with organic nutrients, the 10-ml volume of added liquid contained various concentrations (undiluted, and diluted by factors of 100 and 10,000 of beef extract plus peptone (Difco Laboratories, Detroit, Michigan, USA), a standard source of complex organic nutrients for culturing a wide variety of microorganisms. Net preferences, which indicated the extent to which soil (or soil + organic nutrients) preferentially degraded an enantiomer, were absolute values of differences in percentage of samples preferentially degrading (+) or (–) enantiomers. All dilutions of the organic nutrients affected enantioselectivity similarly; therefore, these results were pooled in Table 2. Because Brazilian forest and pasture samples exhibited almost identical degrees of preference for methyl dichlorprop enantiomers, results from these samples were also combined in Table 1.

To operationally determine the taxonomic relationships between bacterial isolates, restriction banding patterns of the 16S rDNA of each isolate were compared. Phenol/chloroform-extracted DNA from each bacterial isolate was used as the template in a polymerase chain reaction (PCR) to amplify 16S rDNA. The primers used in the PCR (27f (5'-AGAGTTTGTATCTGGCTCAG-3') and 1492r (5'-GGTACCTTGTTACG ACTT-3')) yielded an amplicon corresponding to bases 8 through 1492 (*Escherichia coli* numbering) of the bacterial 16S rDNA. After the PCR, the 16S rDNA amplicon was purified by phenol/chloroform extraction and digested with two restriction enzymes, *RsaI* and *AluI*. Restriction products were separated by electrophoresis on a highly resolving 3% agarose gel (Metaphor, FMC, Rockland, Maine, USA) and visualized using a laser gel scanner and the double-stranded DNA fluorescent stain SYBR Gold (Molecular Probes, Portland, Oregon, USA). Gel lanes were normalized to a single molecular weight standard and analysed using the computer software program Molecular Analyst. Banding patterns were compared based on constituent band positions. The cluster diagram, based on differences in ARDRA banding patterns, was produced by UPGMA (unweighted pair group method using arithmetic averages) of a Jaccard similarity matrix. The cophenetic correlation was 77%. To investigate the reliability of ARDRA relationships a portion (~720 base pairs) of the 16S rDNA was sequenced from the isolates shown in bold-faced type in Fig. 2. In general, isolates with similar or identical banding patterns also had similar or identical 16S rDNA sequences.

Mortality rates with ruelene were determined by exposing North American forest tent caterpillars (*Malacosoma disstria*). Forty larvae from a single nest were divided into two equal groups and fed equal amounts of native black cherry leaves coated with either the (+) or (–) enantiomer. The insecticide was dissolved in acetone (200 mg ml⁻¹), thinly spread over the surface of the leaves and allowed to air dry. Similar tests, in which plantains (*Plantago* sp.) were treated with 700 mg l⁻¹ of separate enantiomers of methyl dichlorprop, showed that, as with dichlorprop, only the (+) enantiomer is herbicidal.

Received 18 May; accepted 14 September 1999.

1. Falconer, R. L., Bidleman, T. F., Gregor, D. J., Semkin, R. & Teixeira, C. Enantioselective breakdown of α -hexachlorocyclohexane in a small arctic lake and its watershed. *Environ. Sci. Technol.* **29**, 1297–1302 (1995).
2. Jantunen, L. M. & Bidleman, T. F. Air-water gas exchange of hexachlorocyclohexanes (HCHs) and the enantiomers of α -HCH in arctic regions. *Geophys. Res.* **108**, 28837–28846 (1996).
3. Aigner, E. J., Leone, A. D. & Falconer, R. I. Concentrations and enantiomeric ratios of organochlorine pesticides in soils from the U.S. corn belt. *Environ. Sci. Technol.* **32**, 1162–1168 (1998).
4. Wiberg, K. et al. Enantioselective gas chromatography/mass spectrometry of methylsulfonyl PCBs with application to arctic marine mammals. *Anal. Chem.* **70**, 3845–3852 (1998).
5. Garrison, A. W., Schmitt, P., Martens, D. & Ketttrup, A. Enantiomeric selectivity in the environmental degradation of dichlorprop as determined by high-performance capillary electrophoresis. *Environ. Sci. Technol.* **30**, 2449–2455 (1996).
6. Tett, V. A., Willetts, A. J. & Lappin-Scott, H. M. Enantioselective degradation of the herbicide mecoprop [2-(2-methyl-4-chlorophenoxy) propionic acid] by mixed and pure bacterial cultures. *Microbial Ecol.* **14**, 191–200 (1994).
7. Zipper, C., Suter, M. J.-F., Haderlein, M. G. & Kohler, H.-P. E. Changes in the enantiomeric ratio of (R) to (S)-mecoprop indicate in situ biodegradation of this chiral herbicide in a polluted aquifer. *Environ. Sci. Technol.* **32**, 2070–2076 (1998).
8. Kohler, H.-P. E. et al. Environmental fate of chiral pollutants—the necessity of considering stereochemistry. *Chimia* **51**, 947–95 (1997).
9. Williams, A. Opportunities for chiral chemicals. *Pestic. Sci.* **46**, 3–9 (1996).
10. Counting on Chiral Drugs. *Chem. Eng. News* **76**(38), 83 (1998).
11. Buser, H. P. & Francotte, E. in *Chiral Separations* (ed. Ahuja, S.) Ch. 5 (American Chemical Society, Washington DC, 1997).
12. Peterjohn, W. T. et al. Responses of trace gas fluxes and N availability to experimentally elevated soil temperatures. *Ecol. Appl.* **4**, 617–625 (1994).
13. Lukeville, A. & Wright, R. F. Experimentally increased soil temperature causes release of nitrogen at a boreal forest catchment in southern Norway. *Glob. Change Biol.* **3**, 13–21 (1997).
14. Hollibaugh, J. T. Metabolic adaptation in natural bacterial populations supplemented with amino acids. *Estuar. Coast. Mar. Sci.* **9**, 215–230 (1979).
15. Akkermans, A. D. L., van Elsas, J. D. & de Bruijn, F. J. *Molecular Microbial Ecology Manual* 1–488 (Kluwer Academic, Dordrecht, 1995).
16. Fulthorpe, R. R., Rhodes, A. N. & Tiedje, J. M. High levels of endemicity of 3-chlorobenzoate-degrading soil bacteria. *Appl. Environ. Microbiol.* **64**, 1620–1627 (1998).
17. Moyer, C. L., Dobbs, F. C. & Karl, D. M. Estimation of diversity and community structure through restriction fragment length polymorphism distribution analysis of bacterial 16S rRNA genes from a microbial mat at an active, hydrothermal vent system. Loihi Seamount, Hawaii. *Appl. Environ. Microbiol.* **60**, 871–879 (1994).
18. Müller, M. D. & Buser, H.-R. Environmental behavior of acetamide pesticide stereoisomers. 2. Stereo- and enantioselective degradation in sewage sludge and soil. *Environ. Sci. Technol.* **29**, 2031–2037 (1995).
19. Standards for the use or disposal of sewage sludge. In *Code of Federal Regulations*, Title 40, Part 503, Subpart B, *Land Application* (US Government Printing Office, revised 1 July 1998).
20. Lewis, D. L. EPA science: casualty of election politics. *Nature* **381**, 731–732 (1996).

Acknowledgements

We thank R. Wright and C. Cerri for collecting soil samples, R. E. Hodson and J. T. Hollibaugh and S. Karickhoff for suggestions, and J. Avants, G. Chapman, and L. Howell for performing chiral analyses. J. Maudsley and D. Gattie assisted with bioassays. K.E.W. was supported by a National Research Council post-doctoral fellowship.

Correspondence and requests for materials should be addressed to D.L.L. (e-mail: LewisDaveL@aol.com).

Mixing and convection in the Greenland Sea from a tracer-release experiment

A. J. Watson*, M.-J. Messias*, E. Fogelqvist†, K. A. Van Scoy‡, T. Johannessen§, K. I. C. Oliver*, D. P. Stevens||, F. Rey¶, T. Tanhua†, K. A. Olsson†, F. Carse*, K. Simonsen#, J. R. Ledwell*, E. Jansen§, D. J. Cooper‡, J. A. Kruepke‡ & E. Guilyardi*

* School of Environmental Sciences, University of East Anglia, Norwich NR4 7TJ, UK

† Department of Analytical and Marine Chemistry, Göteborg University, S-41296, Göteborg, Sweden

‡ Atmospheric and Oceanic Sciences, University of Wisconsin-Madison, 1225 West Dayton Street, Madison, Wisconsin 53706, USA

§ University of Bergen, Department of Geology, Allégaten 71, N-5007 Bergen, Norway

|| School of Mathematics, University of East Anglia, Norwich NR4 7TJ, UK

¶ Institute of Marine Research, Department of Marine Environment, Postboks 1870-Nordnes, N-5024 Bergen, Norway

The University of the Faroe Islands, Noatun, FO-100, Tórshavn, Faroe Islands

* Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA

Convective vertical mixing in restricted areas of the subpolar oceans, such as the Greenland Sea, is thought to be the process responsible for forming much of the dense water of the ocean interior^{1,2}. Deep-water formation varies substantially on annual and decadal timescales^{3–5}, and responds to regional climate signals such as the North Atlantic Oscillation^{6,7}; its variations may therefore give early warning of changes in the thermohaline circulation that may accompany climate change⁸. Here we report direct measurements of vertical mixing, by convection and by turbulence, from a sulphur hexafluoride tracer-release experiment in the central Greenland Sea gyre. In summer, we found rapid turbulent vertical mixing of about $1.1 \text{ cm}^2 \text{ s}^{-1}$. In the following late winter, part of the water column was mixed more vigorously by convection, indicated by the rising and vertical redistribution of the tracer patch in the centre of the gyre. At the same time, mixing outside the gyre centre was only slightly greater than in summer. The results suggest that about 10% of the water in the gyre centre was vertically transported in convective plumes, which reached from the surface to, at their deepest, 1,200–1,400 m. Convection was limited to a very restricted area, however, and smaller volumes of water were transported to depth than previously estimated⁹. Our results imply that it may be the rapid year-round turbulent mixing, rather than convection, that dominates vertical mixing in the region as a whole.

The experiment was begun by releasing, in streaks, 320 kg of sulphur hexafluoride (SF_6) in the central Greenland Sea during August 1996 (Fig. 1). By continually adjusting the depth of the tracer injection apparatus, the tracer was injected as nearly as possible on a single density surface (potential density referenced to 500 dbar, $\sigma_{0.5} = 30.4268$, with standard deviation 0.0004) near 300 m depth. This enabled its subsequent spread to be used to accurately measure diapycnal (cross-density) mixing^{10–12}. The region is dominated by a cyclonic circulation², steered by the bottom topography, and the resulting upward-doming of the density structure in the centre of the gyre is believed to help “precondition” the water column there to deep convection^{1,13,14}, as possibly does brine release into the water during ice formation in winter¹⁵.

Following the release of the tracer, its distribution was documented in November 1996, February–March 1997 and April–May 1997, the analytical technique used being that described by Law *et al.*¹⁶.

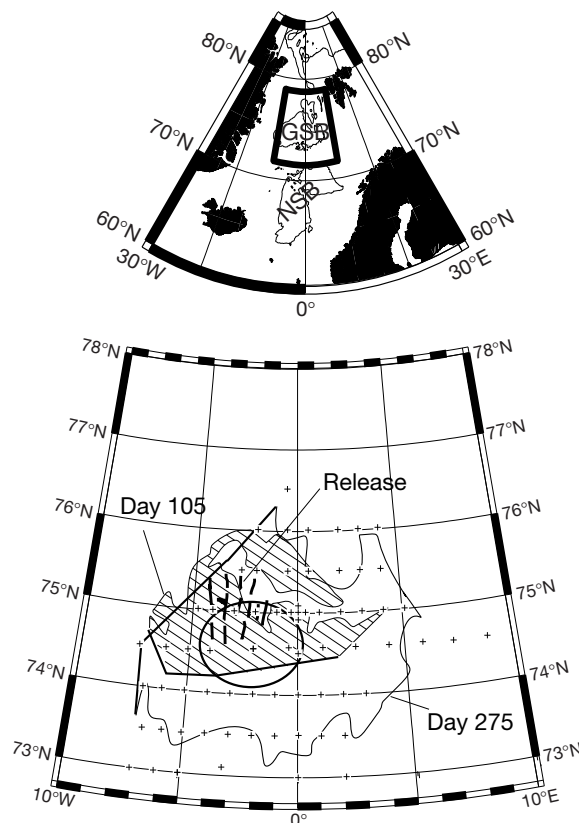


Figure 1 The location of the experiment. The upper map shows the Nordic seas, with the 3,000-m depth contour drawn to delineate the Greenland Sea basin (GSB) and the Norwegian Sea basin (NSB). The region in the box is shown in the lower map. Thick lines show streaks of tracer released in August 1996. Thinner lines show the position of the patch (defined as the region within the 10 nmol m^{-2} contour of SF_6 column integral) found on surveys at about 105 days after release (survey 1; hatched area) and 275 days (survey 3). Thicker edges show where the edge of the tracer patch could not be delineated because of the presence of ice. The ellipse encloses the stations considered as the ‘gyre centre’. Crosses show positions of stations during survey 3.

Each survey consisted of more than 50 stations at which the depth distribution of tracer was documented. Representative distributions of the tracer are shown in Fig. 1. Ice conditions were particularly severe in the Greenland Sea during the 1996/1997 winter, and the survey vessel was sometimes unable to delineate the tracer patch to the west and south. Nevertheless, integration of all the samples suggests that on each survey, about 60–70% of the released tracer was accounted for within the sampled area.

Average tracer-versus-depth profiles for the three surveys are shown in Fig. 2. A correction for background concentrations has been made, based on a pre-release survey of concentrations. Mean vertical profiles could be obtained by simply averaging concentrations with depth in a given region, but in this case station-to-station variations in the vertical structure due to internal waves and sloping isopycnals tend to broaden the averages. For this reason, in Fig. 2 we have instead averaged together profiles using $\sigma_{0.5}$ as a vertical ordinate (see Supplementary Information), and used average density versus depth relations calculated from the same profiles to map these into depth^{11,17}. The tracer sank in the water column from survey 1 to survey 2 (see Table 1 for details of survey times), in response to a deepening of the density structure in the survey region, as shown by the arrows in Fig. 2 which indicate the depth of the “target” isopycnal on each survey. On survey 3, after the deep convection period, the peak of the tracer distribution had risen

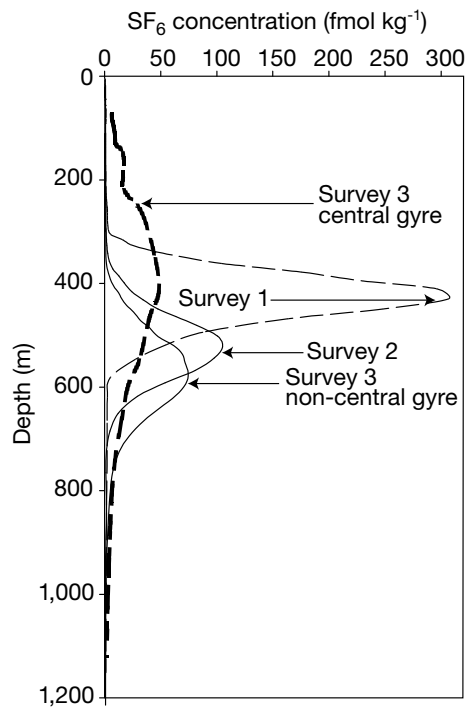


Figure 2 Mean profiles of tracer concentration against depth. Profiles for each of three surveys are shown. For the last survey, the profiles in the centre of the gyre have been plotted separately from those surrounding it. The positions of horizontal arrows show the mean depth at which the potential density corresponding to the tracer release, $\sigma_{0.5} = 30.4268 \pm 0.0004$, was found on each survey.

~150 m in the gyre centre relative to its depth in the same region on survey 2, and had spread extensively vertically. Outside the gyre centre region, these effects were not apparent.

Table 1 gives the vertical mixing coefficients which would be necessary to produce the observed vertical spreading, assuming one-dimensional turbulent mixing, constant in both time and space, occurring between each pair of surveys. The range of values is calculated from two extreme contrasting assumptions about how the depth–density structure changed: namely, that the structure measured on the initial survey applies to the whole period, or that that on the final survey applies to the whole period¹⁷. For the summer and early winter at all locations, these estimates are reasonably robust to the different assumptions, and fall in the range $0.66\text{--}1.13\text{ cm}^2\text{ s}^{-1}$. In spring outside the gyre the vertical mixing increases somewhat, to $2.1\text{--}2.5\text{ cm}^2\text{ s}^{-1}$. In the central gyre post-convection, the stratification had decreased greatly compared to the other surveys, and the mixing rate is not therefore well-defined by this model. The high values, even when calculated using pre-convection depth–density relations, suggest that turbulent exchange was enhanced here, but a more appropriate model is required to interpret the data quantitatively for this period.

From the values in Table 1, we calculate a mean of $1.3 \pm 0.5\text{ cm}^2\text{ s}^{-1}$ for the summer-time, and outside-gyre winter-time, vertical diffusivity. This is substantially larger than the value of $0.12\text{--}0.17\text{ cm}^2\text{ s}^{-1}$ found at a similar depth in the subtropical gyre of the North Atlantic by a recent tracer-release experiment^{12,17}. However, the hydrography of the two sites is very different, and in particular the stratification as measured by the buoyancy frequency N in the present experiment was $0.3\text{--}0.5$ cycles per hour (c.p.h.), compared to ~ 2.5 c.p.h. in the subtropical North Atlantic.

Substantial changes in the temperature and salinity distribution in the gyre centre accompany the evolution of the tracer patch between surveys 2 and 3. The inventories of salt and tracer in the

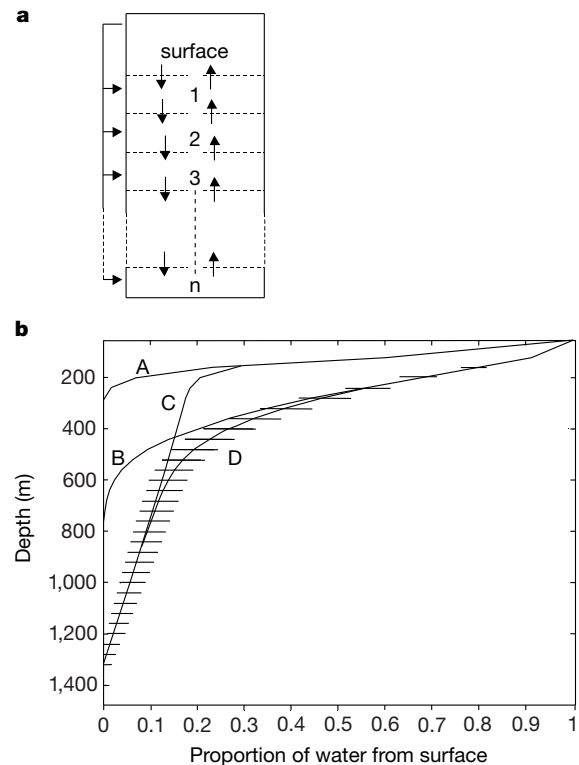


Figure 3 Diagram of the mixing calculation and results. **a**, Diagram of the one-dimensional model for the gyre centre during late winter period. Density of water at the surface can cause water from the surface box to be transferred to depth in the water column, while vertical turbulent exchange connects adjacent boxes (see Methods for details). **b**, Proportion of water mixed from the surface to depth by different processes in the scheme of **a**. Trace A, uniform ‘background’ turbulent mixing ($3\text{ cm}^2\text{ s}^{-1}$) only. Trace B, enhanced turbulent mixing, rising from $3\text{ cm}^2\text{ s}^{-1}$ at $1,360\text{ m}$ to $18\text{ cm}^2\text{ s}^{-1}$ at 200 m . Trace C, surface-to-deep transport by deep convection, increasing linearly from zero at $1,360\text{ m}$ to 20% at 200 m . Trace D, the sum of traces B and C—the model ‘best fit’ to the temperature, salinity and SF_6 data. Error bars show 99% confidence limits.

gyre centre remain nearly constant but are redistributed vertically, consistent with the idea that changes are due mostly to vertical processes rather than horizontal exchange. We have therefore applied a one-dimensional model, shown in Fig. 3a and described in some detail in the Methods section, to the gyre centre between surveys 2 and 3. Figure 3b shows model results for the amount of water at each depth deriving from the surface at the time of survey 3. Error bars show 99% confidence levels. Best fits were obtained with deep convection involving $20 \pm 2\%$ of the water column at 200 m depth, decreasing linearly to zero at $1,300 \pm 100\text{ m}$ depth.

The overall effect of convection on vertical mixing during the period August 1996–May 1997 was a modest increase in ‘effective mixing’ in the gyre centre compared to the surrounding region (Table 1, estimates of k_z from release). Because the convection took place over a restricted area ($2\text{--}4 \times 10^{10}\text{ m}^2$, which is $10\text{--}20\%$ of the area of the Greenland Sea) the actual volume of water involved is rather small ($2\text{--}4 \times 10^{12}\text{ m}^3$), and the effect was relatively quickly diluted into the surrounding waters during the following year. Convection was not sufficiently widespread or intense to homogenize the centre of the gyre, so that a pronounced tracer peak remained in the vertical, as well as substantial horizontal inhomogeneities in concentration. The small-scale turbulent vertical mixing, relatively vigorous here compared to other parts of the ocean, thus dominated the vertical redistribution of water properties over the Greenland Sea as a whole. Studies which assume that vertical spreading of tracer properties in the water column is

Table 1 Coefficients of vertical turbulent exchange

	Mean time from release (d)	k_z from release ($\text{cm}^2 \text{s}^{-1}$)	k_z from survey 1 ($\text{cm}^2 \text{s}^{-1}$)	k_z from survey 2 ($\text{cm}^2 \text{s}^{-1}$)
Survey 1	101	1.11–1.14		
Survey 2	199	0.85–1.1.13	0.66–0.84	
Survey 3 (non-central region)	274	1.03–1.35	1.39–1.90	2.16–2.49
Survey 3 (central region)	274	1.49–6.03	1.07–7.30	3.46–14.87

These coefficients (k_z) were calculated from the data in Fig. 2; using the relation $2k_z t = s_2^2 - s_1^2$, where t is the time separating two profiles of second moment s_1^2 and s_2^2 . Second moments were found by least-squares fits of gaussian curves to the mean tracer–depth profiles. Profiles were averaged using potential density ($\sigma_{\theta,s}$) as a vertical ordinate. The lower and upper estimates for each rate correspond to the use of the initial and final mean depth–versus–density profile, respectively (see Supplementary Information). Consistent estimates are obtained except, as shown in the last row of the table, in the central gyre over the period during which convection occurs.

dominated by convection may therefore over-estimate its importance. For example, our estimate for the volume of water involved in convection is about an order of magnitude less than that derived by Rhein⁹ from a study of chlorofluorocarbons in the Greenland Sea. Historically, 1996/1997 was a moderately active year for convection; in some of the early years of this decade, convection was confined to within a few hundred metres of the surface, while in the period 1960–80 it is believed to have penetrated to great depth resulting in bottom water renewal¹⁸. Whereas those active years left a signal in local water-mass properties which persists for decades, the signal due to the moderate convection of 1996/1997 is likely to be quickly erased.

Apart from the measurement of the effects of mixing in the gyre, a second important purpose of the tracer-release experiment is to test alternative hypotheses for the role of the Greenland Sea as a source for deep waters of the North Atlantic^{19,2} and Norwegian Sea Deep Waters²⁰. For these reasons, concentrations of SF₆ continue to be monitored in the region. As of summer 1998, tracer was undetectable in any of the overflows to the North Atlantic, but tracer had been observed moving into the Norwegian Sea basin, and also spreading northward into the Arctic Ocean via the Fram Strait. □

Methods

One-dimensional model

This numerical scheme was used to diagnose convection and mixing. Figure 3a shows the transports included in the calculation. The water column in the central gyre is represented by a stack of interconnected boxes for each of which the average temperature, salinity and tracer concentration are specified at the time of surveys 2 and 3. Vertically integrated changes of heat, salt and tracer content then give the surface exchanges of heat, fresh water, and tracer (though the last is negligible). Heat and fresh water are exchanged only from the uppermost box, but turbulent exchange and convection of dense plumes carrying cold surface water into the interior may cool the deeper boxes. The convected volume is assumed to decrease linearly with depth, reflecting a decrease in number of plumes able to penetrate into the water column with increasing depth. The volume of convecting water as a function of depth is then specified by the volume leaving the first box and the rate of linear decrease with depth, both of which are considered as free parameters. To conserve volume, a compensatory upwelling, increasing towards the surface, is assumed. Convection is assumed also to enhance turbulent diffusion, the enhancement increasing linearly from zero at the maximum depth of convection and its rate of increase being also treated as a free parameter. The equations were initialized with mean profiles from the centre of the gyre on survey 2, and stepped forward to simulate the evolution of the gyre centre up to survey 3. Over many runs, the free parameters were varied over a range of values. The run having the 'best-fit' values of these parameters was determined by minimizing the following χ^2 parameter²¹:

$$\sum_{j=1}^3 \sum_{i=2}^n \frac{(x_{ij} - X_{ij})^2}{\sigma_{ij}^2}$$

where $j = 1, 2, 3$ specifies potential temperature, salinity or SF₆, and $i = 2, \dots, n$ enumerates the levels (below the surface box) in the model, x_{ij} is the model output and X_{ij} the corresponding measurements on survey 3. The uncertainties σ were estimated for potential temperature, salinity and SF₆ by calculating the error on the mean of these quantities at each depth in the gyre-centre profiles of survey 3.

Our observations do not tell us at what time between the surveys the convection occurs. The time sequence assumed for the convection has a small influence on the best-fit

predictions. For example, if all the convection is assumed to occur in a single burst at the beginning of the modelled period, the heat, salt and tracer transported by the convection would not be identical to the case where the same total amount of water convects to the same depths, but at a uniform rate spread over the entire period. Different assumptions about the timing of convection were tested using heat and moisture fluxes derived for 6-h timesteps from the European Centre for Medium Range Weather Forecast (ECMWF) model. It was assumed that convection was active only when the heat loss rose above some prescribed threshold value. The integrated heat loss over the period from the ECMWF was smaller than that observed in the water column, and the ECMWF heat fluxes were therefore re-scaled to match the observed heat loss. The results shown in Fig. 3b are plotted with a threshold value of 380 W m⁻², which corresponds to 4.25 d of convection and cooling of the deeper layers over the 75-d period between the surveys. However, the results were extremely insensitive to the duration of convection; for threshold values between 0 and 500 W m⁻² (corresponding to 65 to 0.5 d of convection), the best-fit maximum depth of convection varied <100 m and the total amount of water involved varied <10%.

The model corresponds to the situation in which a population of convective plumes initiated by surface cooling penetrate to various depths, with the number of plumes decreasing as the depth of penetration increases, down to some maximum penetration. Individual plumes have a duration of hours to days only²², and contain densified surface water and subsurface water, efficiently mixed²³, which subsequently spreads out at the resulting density level in the water column. Injection of water to levels at and below the main tracer peak caused the tracer distribution to rise and spread vertically, and these characteristics of the measurements therefore place strong constraints on the total amount of water involved in the convection process and the depths to which it is transported.

Received 25 June; accepted 6 August 1999.

1. Killworth, P. D. Deep convection in the world ocean. *Rev. Geophys. Space Phys.* **21**, 1–21 (1983).
2. Rudels, B. The thermohaline circulation of the Arctic Ocean and the Greenland Sea. *Phil. Trans. R. Soc. Lond. A* **352**, 287–299 (1995).
3. Mosby, H. Deep Water in the Norwegian Sea. *Geofis. Publ.* **21**(3), 1–62 (1959).
4. Schlosser, P., Bönsch, G., Rhein, M. & Bayer, R. Reduction of deep-water formation in the Greenland Sea during the 1980s—evidence from tracer data. *Science* **251**, 1054–1056 (1991).
5. Meincke, J., Jonsson, J. & Swift, J. H. Variability of convective conditions in the Greenland Sea. *Int. Council Explorat. Seas Mar. Sci. Symp.* **195**, 32–39 (1992).
6. Dickson, R., Lazier, J., Meincke, J., Rhines, P. & Swift, J. Long term co-ordinated changes in the convective activity of the North Atlantic. *Prog. Oceanogr.* **38**, 241–295 (1996).
7. Delworth, T. L., Manabe, S. & Stouffer, R. J. Multidecadal climate variability in the Greenland Sea and surrounding regions: A coupled model simulation. *Geophys. Res. Lett.* **24**, 257–260 (1997).
8. Manabe, S. & Stouffer, R. J. Century-scale effects of increased atmospheric CO₂ on the ocean–atmosphere system. *Nature* **364**, 215–218 (1993).
9. Rhein, M. Convection in the Greenland Sea, 1982–1993. *J. Geophys. Res.* **101**, 18183–18192 (1988).
10. Watson, A. J. & Ledwell, J. R. Purposefully released tracers. *Phil. Trans. R. Soc. Lond. A* **325**, 189–200 (1988).
11. Ledwell, J. R. & Watson, A. J. The Santa-Monica Basin tracer experiment—a study of diapycnal and isopycnal mixing. *J. Geophys. Res.* **96**, 8695–8718 (1991).
12. Ledwell, J. R., Watson, A. J. & Law, C. S. Evidence for slow mixing across the pycnocline from an open-ocean tracer-release experiment. *Nature* **364**, 701–703 (1993).
13. Killworth, P. D. On "chimney" formations in the ocean. *J. Phys. Oceanogr.* **9**, 531–554 (1979).
14. Ivey, G. N., Taylor, J. R. & Coates, M. J. Convectively driven mixed-layer growth in a rotating, stratified fluid. *Deep Sea Res.* **142**, 331–349 (1995).
15. Visbeck, M., Fischer, J. & Schott, F. Preconditioning the Greenland Sea for deep convection—ice formation and ice drift. *J. Geophys. Res.* **100**, 18489–18502 (1995).
16. Law, C. S., Watson, A. J. & Liddicoat, M. I. Automated vacuum analysis of sulfur-hexafluoride in seawater—derivation of the atmospheric trend (1970–1993) and potential as a transient tracer. *Mar. Chem.* **48**, 57–69 (1994).
17. Ledwell, J. R., Watson, A. J. & Law, C. S. Mixing of a tracer in the pycnocline. *J. Geophys. Res.* **103**, 21499–22529 (1998).
18. Bönsch, G., Blindheim, J., Bullister, J. L., Schlosser, P. & Wallace, D. W. R. Long-term trends of temperature, salinity, density, and transient tracers in the central Greenland Sea. *J. Geophys. Res.* **102**, 18553–18571 (1997).
19. Mauritzen, C. Production of dense overflow waters feeding the North-Atlantic across the Greenland-Scotland Ridge. 1. Evidence for a revised circulation scheme. *Deep Sea Res.* **143**, 769–806 (1996).
20. Aagaard, K., Fahrbach, E., Meincke, J. & Swift, J. H. Saline outflow from the Arctic-Ocean—its contribution to the deep waters of the Greenland, Norwegian, and Iceland Seas. *J. Geophys. Res.* **96**, 20433–20441 (1991).
21. Bevington, P. R. & Robinson, K. *Data Reduction and Error Analysis for the Physical Sciences* (McGraw-Hill, New York, 1992).
22. Klinger, B. A., Marshall, J. & Send, U. Representation of convective plumes by vertical adjustment. *J. Geophys. Res.* **101**, 18175–18182 (1996).
23. Send, U. & Marshall, J. Integral effects of deep convection. *J. Phys. Oceanogr.* **25**, 855–872 (1995).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank the staff of RV *Håkon Mosby*, RV *Johann Hjort* and RRS *James Clark Ross* for their support, and B. Guest, S. C. Sutherland, M. I. Liddicoat, R. D. Ling and T. Fileman for assistance. The main support for this work was from the EU MAST-III programme, European sub-polar ocean project, phase 2. Additional support from the following national agencies was also important: NERC (UK) NRC (Norway) and NSF (USA).

Correspondence and requests for materials should be addressed to A.J.W. (e-mail: a.watson@uea.ac.uk).

Ecosystem consequences of wolf behavioural response to climate

Eric Post*, Rolf O. Peterson†, Nils Chr. Stenseth* & Brian E. McLaren‡

* Department of Biology, Division of Zoology, University of Oslo, PO Box 1050, Blindern, N-0316 Oslo, Norway

† School of Forestry and Wood Products, Michigan Technological University, Houghton, Michigan 49931, USA

‡ Department of Forest Resources, Wildlife Division, Building 810, Pleasantville, St. John's, Newfoundland, A1B 4J6 Canada

Because apex predators exert considerable influence on the structure and function of top-down ecosystems^{1–3}, their responses to climate may shape responses at lower trophic levels⁴. Previous reports of trophic cascades and ecosystem dynamics induced by predators have focused on changes in their abundance^{5–8}, whereas we investigated whether changes in predator behaviour could precipitate cascades of similar ecological scale. Here we report the ecological consequences of predator behavioural response to global climatic variation using 40 years of data on wolf predation from Isle Royale, USA, where wolves limit abundance of moose⁹, which limit productivity of fir trees¹⁰. In response to increases in winter snow related to the North Atlantic Oscillation, wolves hunted in larger packs and, consequently, tripled the number of moose killed per day compared with less snowy years when they hunted in smaller packs. Following increased predation rates, moose abundance declined, and, following release from heavy browsing, growth of understory fir increased. Hence, cascading

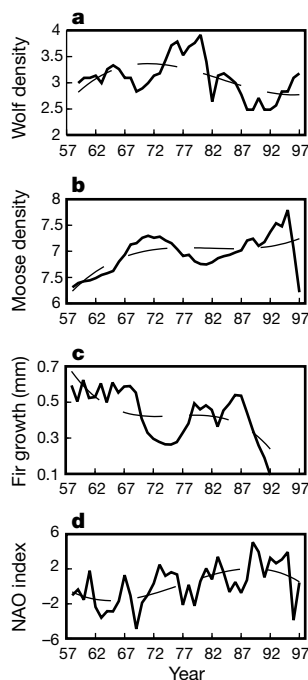


Figure 1 Dynamics of the three-trophic-level system on Isle Royale, USA, and the North Atlantic Oscillation index, 1958–97. **a**, Interannual variation in wolf abundance (ln-transformed). **b**, Interannual variation in moose abundance (ln-transformed). **c**, Interannual variation in the mean annual growth increment (in millimetres) of balsam fir trees. **d**, The winter (December–March) North Atlantic Oscillation (NAO) index. The NAO index¹⁷ correlates negatively with snow accumulation on Isle Royale during the current winter and up to six previous winters⁹. Dashed lines are multi-annual trends estimated with third-order polynomial regression. Density is measured in number of animals over the area studied.

behavioural responses of apex predators may be a substantial link in the pathway from climatic change to ecosystem function.

Isle Royale, USA, is a protected national park and the focus of the longest continuous study of an undisturbed, three-trophic-level system involving grey wolves (*Canis lupus*), moose (*Alces alces*) and their primary winter forage, balsam fir (*Abies balsamea*) (Fig. 1). Moose play a pivotal role in ecosystem function on Isle Royale because they are the primary prey of wolves^{11,12}, and because their heavy browsing on balsam fir and other woody species determines fir seedling establishment¹³, sapling recruitment¹⁴ and growth rates¹⁰, forest litter production¹⁵ and edaphic nutrient dynamics¹⁶. Recently, we reported that large-scale climatic variability influences wolf–moose dynamics in this system⁹ through interannual and decadal fluctuations in snow depth related to the North Atlantic Oscillation (NAO)¹⁷ (Fig. 1). Here we present the hypothesis that wolf predatory behaviour links climate and ecosystem dynamics on Isle Royale through wolf–moose and moose–fir dynamics.

Wolf-pack dynamics and hunting behaviour on Isle Royale are recorded systematically during winter aerial surveys initiated in 1959, shortly after wolves colonized the island. Variation in the annual mean size of packs between 1959 and 1998 showed a close correlation with the state of the NAO (Fig. 2a), which correlated negatively with winter snow depth⁹ on Isle Royale. In turn, as pack size increased in response to winter snow depth, the number of moose killed per pack per day rose during the same period (Fig. 2b). Hence, moose winter mortality was strongly dependent on wolf-pack size, which related directly to large-scale variation in winter climate.

Examination of carcasses of moose killed by wolves during winter revealed that as pack size increased, wolves also killed more calves (correlation coefficient $r = 0.52$, significance level $P = 0.01$) and old moose ($r = 0.55$, $P = 0.05$). For old moose, this reflected both an increase in pack size with snow depth and a direct influence of climate on the susceptibility of old moose to predation, as the partial correlation between old moose killed and the NAO index ($r = -0.52$, $P < 0.05$) was independent of pack size. Additionally,

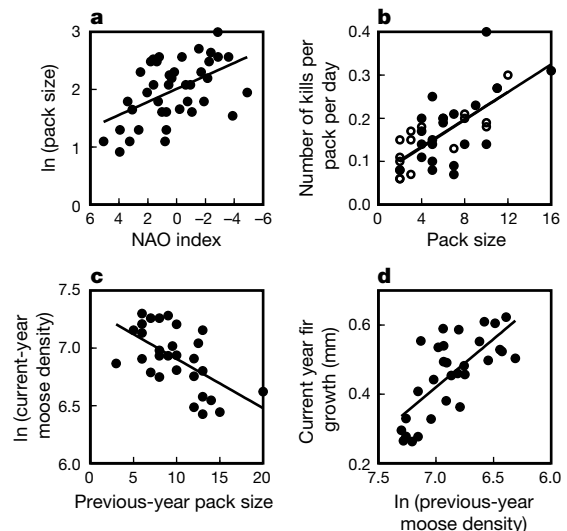


Figure 2 Progression of climatic influence on ecosystem function from wolf behaviour to growth of fir trees. **a**, Increase in the mean size of wolf packs during snowy (negative NAO) winters, 1959–98 ($r = -0.51$, $P < 0.02$). **b**, Increase in the winter kill rate of wolf packs with pack size, 1972–98, for the East Pack (solid circles, $R^2 = 0.37$, $P < 0.01$) and West Pack (open circles, $R^2 = 0.69$, $P < 0.001$). Trend lines for each pack overlap completely. Common regression: $R^2 = 0.49$, $P < 0.001$. **c**, Decline in moose density one year after increase in size of winter wolf packs, 1959–88 ($r = -0.58$, $P < 0.05$). **d**, Increased growth of fir trees one year after decline in moose density, 1958–88 ($R^2 = 0.52$, $P < 0.001$). Density is measured in number of animals over the area studied; pack size is measured in number of wolves per pack.

in support of earlier observations^{9,18,19}, the condition and survival of adult moose declined after snowy winters. Examination of summer wolf faeces revealed increases in both the per cent occurrence ($r = 0.79$, $P = 0.02$) and biomass ($r = -0.83$, $P = 0.01$) of adult moose following negative (snowy) NAO winters.

Direct observations of numbers of wolf pups, together with estimates of the percentage of calves and numbers of old moose in the population, indicated that wolf productivity was positively related to both moose production and the availability of calves and old moose (Table 1) (total multiple regression coefficient $R^2 = 0.34$, $P = 0.037$). Also, annual mortality of wolves, which was density-dependent, declined following winters in which wolves hunted in large packs (Table 1) (total $R^2 = 0.55$, $P = 0.001$). Hence, availability of calves and old moose determined productivity of wolves²⁰, and winter wolf-pack size influenced survival of wolves.

Because the amount of food acquired per individual wolf declines with increasing pack size^{21,22}, it is not obvious that wolves hunting in larger packs should enjoy both greater productivity and survival. However, analysis of estimates of hunting success per individual wolf²³ revealed that, in addition to the increased kill rate per pack during snowy winters (Fig. 2b), the kill rate per individual wolf also increased during snowy winters (Table 1), indicating that moose were easier to kill during the snowiest winters. Thus, the parallel increase in hunting efficiency with winter snow translated into greater food acquisition by mating pairs and their offspring. The tendency of both moose and wolves to travel along the shorelines of Lake Superior and inland lakes during snowy winters increases the frequency of encounters between the two species²⁴, and, together with greater pack sizes, probably also contributed to the higher kill rates observed during snowy winters.

Climatic influences on the hunting behaviour of wolves apparently cascaded down onto secondary and primary producers in this system. One year after snowy winters, increases in wolf-pack size and killing rates led directly to a decline in moose abundance (Fig. 2c), independently of density-dependent limitation of moose (total $R^2 = 0.92$, $P < 0.001$). Analysis of tree rings from the west end of Isle Royale, where fir are heavily and repeatedly browsed by moose¹⁰, revealed that growth of understory balsam fir saplings increased after snowy winters: the annual growth increment of fir correlated negatively with the NAO index of the current year ($r = -0.59$, $P < 0.005$) and previous year ($r = -0.42$, $P < 0.05$).

However, multiple regression revealed that the one-year-delayed effect of winter climate on growth of fir related directly to release from browsing after declines in moose abundance (Fig. 2d), whereas the current-year effect was related to the NAO (partial $r = -0.48$, $P < 0.05$). This suggests, in contrast to an earlier analysis¹⁰, that climate may directly influence fir growth on the west end of the island by offering protection from browsing under deep snow. In support of this suggestion, analysis of data from the east end of Isle Royale, where fir have escaped heavy browsing by moose¹⁰, revealed that fir growth there did not correlate with the NAO index or moose density (see Methods). Because deep snow hampers both moose foraging and locomotion^{24,25}, the pathway of climatic influences on trophic dynamics in this system may begin with vulnerability of moose, but cascade down from wolves once they begin killing more frequently.

This study indicates that trophic cascades may occur in a three-level system as a result of behavioural responses of an apex predator to large-scale climatic variability. Such behaviourally induced cascades may ultimately influence ecosystem function, as primary production and edaphic nutrient dynamics are strongly influenced by herbivores such as moose^{13–16,26}. Moreover, although pack size in wolves is believed to be limited by the minimum number of individuals needed to find and kill prey and by social competition²⁷, the influence of climate on sociality in wolves illustrates that environment can influence behaviour in species displaying complex social organization, with ecosystem consequences. Similarly, the social behaviour of fish-eating predators links nutrient dynamics

Table 1 Dynamics of wolf production, mortality and hunting success on Isle Royale, USA

Variable (year)	Standardized coefficient	Partial <i>r</i>	<i>F</i>
Per cent pups (<i>t</i>)			
Per cent moose calves (<i>t</i>)	0.61*	0.42*	7.69*
Density of old moose (<i>t</i>)	0.81*	0.40*	7.29*
Total moose density (<i>t</i> – 1)	–	0.45	3.72
Wolf density (<i>t</i> – 1)	–	–0.21	0.67
NAO (<i>t</i> – 1)	–	0.02	0.01
Annual wolf mortality (<i>t</i>)			
Wolf density (<i>t</i> – 1)	0.97*	0.51*	18.34*
Pack size (<i>t</i>)	–0.61*	–0.62*	10.22*
Number of packs (<i>t</i>)	–0.81*	–0.41*	17.86*
NAO (<i>t</i> – 1)	–	–0.15	0.33
Individual kill rate (<i>t</i>)			
NAO (<i>t</i>)	–0.44*	–0.53*	5.66*
Wolf density (<i>t</i>)	–0.38*	–0.66*	4.57*
Pack size (<i>t</i>)	–	0.42	2.95
Moose density (<i>t</i>)	–	–0.36	2.15

Standardized coefficients, partial correlation coefficients and *F*-tests are shown from stepwise multiple linear regression of interannual variation in annual per cent pups (1970–91), annual mortality (1970–91), and number of moose killed per wolf per 100 days during winter (1967–85). Current-year effects are denoted with *t*.

* Significant, $P < 0.05$.

between maritime and terrestrial ecosystems²⁸. Just as the importance of herbivore behaviour (rather than numerical) responses to predators in trophic cascades has gained increasing attention^{29,30}, this study indicates a need for additional focus on links between climate and animal behaviour in ecosystem processes. □

Methods

Wolf dynamics and winter kill data

Abundance of wolves was estimated during aerial surveys between late January and early March, 1959–98, when sizes, numbers and locations of packs were recorded. We used natural-logarithm-transformed estimates of mean annual pack size to remedy heteroscedasticity. Raw data yielded a similar correlation coefficient ($r = -0.47$, $P < 0.01$). The full regression model of wolf pack size ($R^2 = 0.73$, $P < 0.001$) included wolf density (partial $r = 0.66$, $P < 0.02$), number of wolf packs (partial $r = -0.59$, $P < 0.005$) and the NAO index (partial $r = -0.45$, $P < 0.05$). Annual mortality was estimated from absolute changes in wolf abundance together with pup numbers between successive winters²⁰. In analysing pack kill rate versus pack size, we used data from the two most enduring wolf packs, the East Pack and West Pack²⁰. Pack kill rate was not related to moose density or wolf density for either pack.

During winter, wolf travel routes were located from the air and either followed or back-tracked to locate moose kills. Nearly all moose carcasses were examined from the ground to establish age and sex of the kill¹¹. The proportions of calf and old (aged 10–14 years) moose in the annual winter kill were arcsine-transformed before analysis. The annual kill interval during winter for wolf packs (days per kill) was calculated from observations made while back-tracking wolves between kills for 40–60 days each winter.

Per cent occurrence and biomass of adult moose in wolf faeces during summer were estimated from macrohistological analysis of fresh faeces²¹. When regressing per cent occurrence and biomass of adult moose in wolf faeces versus the NAO index, we also accounted for moose density during the previous winter. Partial correlations with moose density were, for per cent occurrence, $r = 0.82$ ($P = 0.01$), and, for per cent biomass, $r = 0.84$ ($P = 0.01$).

Moose dynamics

Annual estimates of abundance of moose and the age structure of the population during 1958–88 were based on cohort reconstruction¹¹ from remains of moose collected from 1958–97. Estimates based on aerial surveys during winter (late January–early March) extend through 1997. Annual per cent calves in the population were estimated from aerial surveys during winter. To analyse the simultaneous influences of wolf-pack size and moose density on variation in moose density (ln-transformed), we used stepwise multiple regression including 'year' to de-trend the data. The first-order autoregressive term quantifying moose density was estimated (± 1 s.e.) as 0.85 ± 0.06 , $P < 0.05$ (a two-tailed *t*-test was used to test for difference from one¹⁹). We analysed effects of wolf-pack size on moose density only for 1959–88 because annual aerial estimates of moose abundance beyond 1988 have not been corrected by cohort reconstruction.

Balsam fir dynamics

We used data on growth of balsam fir saplings from the west end of Isle Royale, where trophic interactions between moose and fir are tightly coupled¹⁰. Fir trees were sampled in 1992, when 8 individuals (aged 48–60 years) were cut for examination of ring-width chronologies¹⁰. The mean annual ring-width index (in millimetres) is based on an average of 4 measurements taken on each tree at 5–10-cm intervals along their stems. In the analysis of the influence of moose density on fir growth, we used data only for the period

1958–88 (see above). Matching data from 8 trees sampled on the east end of Isle Royale, where balsam fir are not heavily or repeatedly browsed by moose¹⁰, provided a natural control for the test of our hypothesis relating climate and wolf behaviour to the dynamics of moose and fir growth. On the east end, fir growth did not correlate with the NAO index at lags of 0 ($r = 0.074$, $P = 0.69$) or 1 year ($r = 0.069$, $P = 0.71$) or with moose density at a lag of 1 year ($r = 0.054$, $P = 0.78$).

The North Atlantic Oscillation

The NAO is a meridional alternation in atmospheric mass balance between pressure centres over the Azores and Iceland. The winter NAO index is calculated on the basis of the normalized sea-level pressure difference between these two centres from December to March¹⁷. A negative difference indicates weak westerly winds across the Atlantic Ocean and unusually warm winters over north eastern North America; opposing conditions prevail during positive years¹⁷. We used data from the Climate Indices website of the National Center for Atmospheric Research, Boulder, USA (<http://www.cgd.ucar.edu/cas/climind/>). Although annual snow depth measures during the study period are not available for Isle Royale, the NAO index correlates negatively with snow depth in nearby Superior National Forest, USA⁹. In all regressions, degrees of freedom for tests of significance of independent variables were adjusted for autocorrelation⁹.

Received 24 March; accepted 11 August 1999.

- Estes, J. A. & Palmisano, J. F. Sea otters: their role in structuring nearshore communities. *Science* **185**, 1058–1060 (1974).
- Spiller, D. A. & Schoener, T. W. A terrestrial field experiment showing the impact of eliminating top predators on foliage damage. *Nature* **347**, 469–472 (1990).
- McPeck, M. A. The consequences of changing the top predator in a food web: a comparative experimental approach. *Ecol. Monogr.* **68**, 1–23 (1998).
- Sanford, E. Regulation of keystone predation by small changes in ocean temperature. *Science* **283**, 2095–2097 (1999).
- Paine, R. T. Food web complexity and species diversity. *Am. Nat.* **100**, 65–75 (1966).
- Paine, R. T. Intertidal community structure: Experimental studies on the relationship between a dominant competitor and its principal predator. *Oecologia* **15**, 93–120 (1974).
- Estes, J. A. & Duggins, D. O. Sea otters and kelp forests in Alaska: generality and variation in a community ecological paradigm. *Ecol. Monogr.* **65**, 75–100 (1995).
- Estes, J. A., Tinker, M. T., Williams, T. M. & Doak, D. F. Killer whale predation on sea otters linking oceanic and nearshore ecosystems. *Science* **282**, 473–476 (1998).
- Post, E. & Stenseth, N. C. Large-scale climatic variability and population dynamics of moose and white-tailed deer. *J. Anim. Ecol.* **67**, 537–543 (1998).
- McLaren, B. E. & Peterson, R. O. Wolves, moose, and tree rings on Isle Royale. *Science* **266**, 1555–1558 (1994).
- Peterson, R. O. *Wolf Ecology and Prey Relationships on Isle Royale* (National Park Services Scientific Ser. No. 11, Washington DC, 1977).
- Peterson, R. O., Page, R. E. & Dodge, K. M. Wolves, moose, and the allometry of population cycles. *Science* **224**, 1350–1352 (1984).
- McLaren, B. E. & Janke, R. A. Seedbed and canopy cover effects on balsam fir seedling establishment in Isle Royale National Park. *Can. J. For. Res.* **26**, 782–793 (1996).
- Brandner, T. A., Peterson, R. O. & Risenhoover, K. L. Balsam fir on Isle Royale: effects of moose herbivory and population density. *Ecology* **71**, 155–164 (1990).
- McInnes, P. F., Naiman, R. J., Pastor, J. & Cohen, Y. Effects of moose browsing on vegetation and litter of the boreal forest, Isle Royale, Michigan, USA. *Ecology* **73**, 2059–2075 (1992).
- Pastor, J., Dewey, B., Naiman, R. J., McInnes, P. F. & Cohen, Y. Moose browsing and soil fertility in the boreal forests of the Isle Royale National Park. *Ecology* **74**, 467–480 (1993).
- Hurrell, J. W. Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation. *Science* **269**, 676–679 (1995).
- Mech, L. D., McRoberts, R. E., Peterson, R. O. & Page, R. E. Relationship of deer and moose populations to previous winters' snow. *J. Anim. Ecol.* **56**, 615–627 (1987).
- Post, E. & Stenseth, N. C. Climatic variability, plant phenology, and northern ungulates. *Ecology* **80**, 1322–1339.
- Peterson, R. O., Thomas, J. N., Thurber, J. M., Vucetich, J. A. & Waite, T. A. Population limitation and the wolves of Isle Royale. *J. Mamm.* **79**, 828–841 (1998).
- Thurber, J. M. & Peterson, R. O. Effects of population density and pack size on the foraging ecology of gray wolves. *J. Mamm.* **74**, 870–889 (1993).
- Schmidt, P. A. & Mech, L. D. Wolf pack size and food acquisition. *Am. Nat.* **150**, 513–517 (1997).
- Peterson, R. O. & Page, R. E. The rise and fall of Isle Royale wolves, 1975–1986. *J. Mamm.* **69**, 89–99 (1988).
- Peterson, R. O. & Allen, D. L. Snow conditions as a parameter in moose-wolf relationships. *Le Naturaliste Canadien* **101**, 481–492 (1974).
- Kelsall, J. P. Structural adaptations of moose and deer for snow. *J. Mamm.* **50**, 302–310 (1969).
- Molvær, E. M. & Bowyer, R. T. Moose herbivory, browse quality, and nutrient cycling in an Alaskan treeline community. *Oecologia* **94**, 472–479 (1993).
- Mech, L. D. *The Wolf* (Univ. Minnesota Press, 1970).
- Ben-David, M., Bowyer, R. T., Duffy, L. K., Roby, D. D. & Schell, D. M. Social behavior and ecosystem processes: river otter latrines and nutrient dynamics of terrestrial vegetation. *Ecology* **79**, 2567–2571 (1998).
- Beckerman, A. P., Uriarte, M. & Schmitz, O. J. Experimental evidence for a behavior-mediated trophic cascade in a terrestrial food chain. *Proc. Natl Acad. Sci. USA* **94**, 10735–10738 (1997).
- Schmitz, O. J., Beckerman, A. P. & O'Brien, K. M. Behaviorally mediated trophic cascades: effects of predation risk on food web interactions. *Ecology* **78**, 1388–1399 (1997).

Acknowledgements

We thank the U.S. National Science Foundation for grants to E.P. and R.O.P., the US National Park Service and Earthwatch for grants to R.O.P. and the Norwegian Science Council (NFR) for a grant to N.C.S. We thank G.-P. Sætre for discussions.

Correspondence and requests for materials should be addressed to E.P. (e-mail: eric.post@bio.uio.no).

Allometric scaling of production and life-history variation in vascular plants

Brian J. Enquist^{*†}, Geoffrey B. West^{†‡}, Eric L. Charnov[§] & James H. Brown^{§†}

^{*} National Center for Ecological Analysis and Synthesis, 735 State Street, Suite 300, Santa Barbara, California 93101-5504, USA

[§] Department of Biology, University of New Mexico, Albuquerque, New Mexico 87131, USA

[†] The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA

[‡] Theoretical Division, T-8, MS B285, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

A prominent feature of comparative life histories is the well documented negative correlation between growth rate and life span^{1,2}. Patterns of resource allocation during growth and reproduction reflect life-history differences between species^{1,2}. This is particularly striking in tropical forests, where tree species can differ greatly in their rates of growth and ages of maturity but still attain similar canopy sizes^{3,4}. Here we provide a theoretical framework for relating life-history variables to rates of production, dM/dt , where M is above-ground mass and t is time. As metabolic rate limits production as an individual grows, $dM/dt \propto M^{3/4}$. Incorporating interspecific variation in resource allocation to wood density, we derive a universal growth law that quantitatively fits data for a large sample of tropical tree species with diverse life histories. Combined with evolutionary life-history theory¹, the growth law also predicts several qualitative features of tree demography and reproduction. This framework also provides a general quantitative answer to why relative growth rate $(1/M)(dM/dt)$ decreases with increasing plant size ($\propto M^{-1/4}$) and how it varies with differing allocation strategies^{5–8}.

Coexistence in diverse ecological communities has been thought to be due, in part, to life-history trade-offs involving allocation of resources^{6,9–13}. There is, however, no generally accepted mechanistic framework for understanding how patterns of allocation influence variation in life histories. Here we show that a general allometric growth model for trees can provide an explanation for much life-history variation. Metabolism produces the energy and materials that are used for all biological processes. A central issue of life history is how over ontogeny the products of metabolism are allocated among maintenance, growth and reproduction. Previous work has shown how the 3/4-power scaling of metabolic rate with body mass in both animals and plants results from physical and biological constraints on the distribution of resources through fractal-like vascular networks^{14–16}.

In Box 1, we derive a production/growth law in terms of above-ground plant mass, M , or basal stem diameter, D . Our data set (Table 1, Fig. 1) consists of a 20-year change in D ($D(20)$) compared with $D(0)$ for 45 species of trees in a tropical dry forest. In total, there were 2,283 individuals, all of reproductive age. Equation (6) predicts that $D^{2/3}(20) = A_i + D^{2/3}(0)$, so that a plot of $D^{2/3}(20)$ versus $D^{2/3}(0)$ for each species should yield a straight line with a universal slope of unity and an intercept, A_i , inversely proportional to wood density, ρ , so that A_i is proportional to ρ^{-1} . We tested this prediction by analysing the data for 45 species (Fig. 2a, Table 1). The average slope of all species, 1.04, is essentially not different from the predicted value of 1.0 (95% confidence intervals (CI): 1.01 to 1.08). Forty of the forty-five species had slopes statistically indistinguishable from the predicted value of 1 (Table 1). There was no systematic trend to deviate above or below 1; furthermore, we expect 5% (≈ 2.5 species) to differ by chance alone at the 0.05 level.

We conclude that production within species scales as $M^{3/4}$. There is, however, considerable interspecific variation in the intercepts. From Box 1, the intercepts should be inversely proportional to wood density, ρ , and this is confirmed by regression analysis ($r = -0.468$, exponent = -0.934 , 95% CI: -1.631 to -0.237 , intercept = 0.6196 , $n = 29$). Weighting each species by multiplying $D^{2/3}$ by ρ collapses the production relationships for each species onto a line with a universal slope of unity and a common intercept (see Fig. 2b and equation (6)). Essentially, there is a trade-off in growth rate (change in stem diameter) with allocation to tissue density: species that allocate less biomass to their stems (light woods) increase in basal diameter faster than species that allocate more to stems (dense woods).

The trees depicted in Fig. 2a are reproducing, so some proportion of production, λ , must be allocated to reproduction rather than growth. The derivation in Box 1 (resulting in equation (7)) and measured empirical reproductive allometries both suggest that λ is independent of size within a species. To obtain the convergence of allometries implied in Fig. 2b requires the stronger constraint that λ be the same across species. Studies reporting relationships between production of reproduction mass and stem diameter indicate that

whereas measured exponents usually approximate to the predicted value of 2, there are differences in normalization constants, and hence in λ , among species^{17–20}. Such variation will not change the predicted slope of unity when data are plotted as in Fig. 2a. If this variation in λ is independent of ρ , however, it will show up as residual variation in plots such as Fig. 2b.

Production data for seven species of temperate trees provide additional evidence that dM/dt is proportional to $M^{3/4}$. Instead of measuring change in diameter, Whittaker and Woodwell²¹ recorded total annual mass production of leaves and twigs, and bark and stem wood as a function of trunk diameter. Relating trunk diameter to annual production for twig and leaf mass, and wood and bark mass, shows that both scale with exponents essentially indistinguishable from the predicted $M^{3/4}$. So mass production in both tropical and temperate trees scales as $M^{3/4}$.

Relative growth rate $(1/M)(dM/dt)$ decreases with increases in individual size and varies across species^{5–8}. Although this decrease has been qualitatively attributed to several mechanisms⁶, a quantitative rule emerges as a direct consequence of equations (2) and (3): $(1/M)(dM/dt) \propto M^{-1/4}$. This rule is predicted to hold even across species growing in the same environment for comparisons when

Box 1

The growth law

As trees continue to increase in size throughout life, and this growth must be fuelled by metabolism, it is reasonable to assume that the growth rate (the rate at which its mass, M , increases over time), is directly proportional to metabolic rate, B , (the rate of gross photosynthesis). Thus, at any time t ,

$$\frac{dM}{dt} = C_G B \quad (2)$$

where C_G is a proportionality constant that can be time dependent. It has been shown that B is proportional to $V^{3/4}$, and stem diameter, D , is proportional to $V^{3/8}$, where V is the total volume of the plant¹⁴. If $\rho = (M/V)$ is the tissue- or species-specific wood density, then, at any time t , these can be expressed as

$$B = C_B \left(\frac{M}{\rho}\right)^{3/4} \quad D = C_D \left(\frac{M}{\rho}\right)^{3/8} \quad (3)$$

where C_B and C_D are corresponding proportionality constants. Implicit in these results is the assumption that the ratio E/ρ , where E is the Young's modulus of elasticity, is constant for all plants¹⁸. Here, we relax the restriction implicit in ref. 14 that ρ is constant and allow it to differ among plant species and to vary with time. Equations (2) and (3) can be combined to give

$$\frac{dD}{dt} = \left(\frac{3C}{2\rho}\right) D^{1/3} \quad (4)$$

where $C \equiv 1/4 C_G C_B C_D^{2/3}$. This can be integrated to give

$$D(t)^{2/3} - D(t_0)^{2/3} = \int_{t_0}^t \frac{C(t)}{\rho(t)} dt \quad (5)$$

where the time dependence of all variables has been made explicit and the integration has been carried out from some initial time, t_0 , up to some arbitrary time, t . Thus, regardless of any possible time dependence of either the proportionality constants or the density, a plot of $D^{2/3}(t)$ versus $D^{2/3}(t_0)$ for fixed times t and t_0 for any species should yield a straight line with a universal slope of unity but with an intercept that depends on the time interval and the species. If, however, ρ of a given plant is taken to be independent of time, but allowed to vary across species, this can be re-expressed as

$$\rho[D(t)^{2/3} - D(t_0)^{2/3}] = \int_{t_0}^t C(t) dt \quad (6)$$

The intercept of the production relationship should therefore be inversely proportional to ρ . If C_G , C_B , and C_D are independent of time then the intercept, namely the righthand side of equation (6), is given by $C(t - t_0)$. Furthermore, if C does not vary among species as implicitly assumed in ref. 14, and the time interval is the same for all species, then weighting $D^{2/3}(t)$ and $D^{2/3}(t_0)$ by ρ should collapse all species onto a universal line of unit slope. The intercept should no longer depend upon the species but only upon the time interval and C (however, see below for the reproductive period).

This framework allows us to recast the mass-production law as a function of basal stem diameter, D ,

$$\frac{dM}{dt} = \left(\frac{C_G C_B}{C_D^2}\right) D^2 \quad (7)$$

This gives the new prediction that across species, dM/dt , for trees of the same diameter is explicitly independent of wood density ρ .

The derivation so far has assumed that the plant is not reproducing, so that all production is given to growth. Growth must, however, slow with the onset of reproduction¹, as some fraction of production, $(\lambda)dM/dt$, is then devoted to reproduction rather than to individual growth, $(1 - \lambda)dM/dt$. The scaling rule for reproductive allocation and growth after the onset of reproduction depends upon the behaviour of λ . Life-history theory^{1,2} predicts the age (size) course of this reproductive allocation. The simplest model is to take λ as a constant both within and across species. A constant λ within species is supported by several studies^{17–20} that have measured allometries for the biomass of reproductive organs, M_R , versus stem diameter, D , in both angiosperms and gymnosperms. M_R reflects the rate of allocation to reproduction. Since published studies show that M_R is generally proportional to D^2 , it follows from equation (7) that M_R is proportional to dM/dt indicating that λ is indeed approximately a constant within species and independent of size. This is an important result, as animals have λ increasing from 0 to 1 over the reproductive period^{1,2}. The fact that trees appear to have nearly constant values of λ probably reflects the constraints of resource harvesting and biomechanics on plant growth and development. Note that if λ is approximately constant, all of the previous growth results are true for all time periods $(t - t_0)$ after the onset of reproduction, with C_G redefined to be $(1 - \lambda)C_G$.

wood density is held constant.

The simplest life-history theory assumes determinate growth: at the onset of reproduction all production ($\lambda = 1$) is given to reproduction and growth ceases¹. If we approximate plant growth

in this way and assume a pre-reproductive production rate $dM/dt = C_G B = C_G C_B (M/\rho)^{3/4}$ as in equations (2) and (3), and if mortality rate is independent of plant size (over a range which includes possible sizes at maturation), well known arguments¹ for

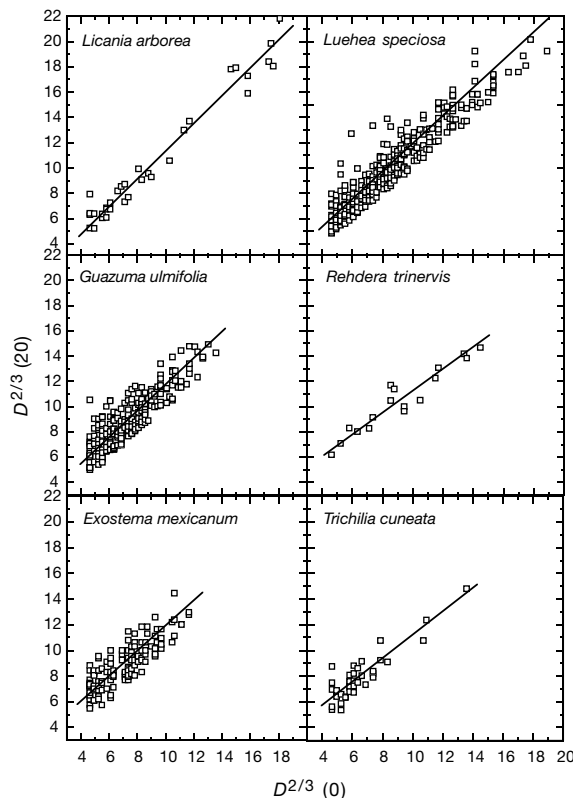


Figure 1 Representative variation of $D^{2/3}(0)$ versus $D^{2/3}(20)$ for six species of tropical trees. When sample sizes are large, the apparent variation is exaggerated because of the overlap of many points. Values for slopes and r^2 for these species are given in Table 1.

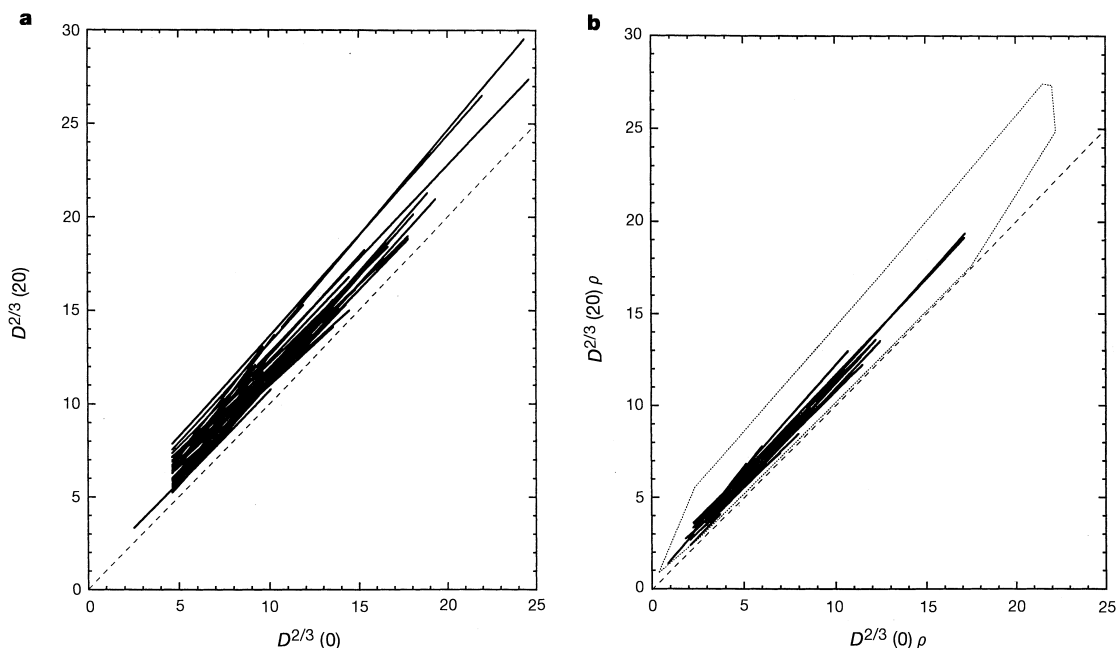


Figure 2 Growth rates of multiple species of tropical forest trees. **a.** $D^{2/3}(0)$ versus $D^{2/3}(20)$ for 45 species of dry forest trees. If production rate scales as $M^{3/4}$ (equation (2)), a slope of 1 (equation (5)) should be obtained. Note that all species have slopes that are essentially indistinguishable from 1, represented by the dashed line of unity. Statistics for each species are reported in Table 1. **b.** Weighting each species production

$D(20)^{2/3} = A_1 + D(0)^{2/3}$ by wood density, ρ , collapses its production relationship onto a common line with unit slope and intercept A (see equation (6)). The dotted box shows the much greater range of variation represented in Fig. 2a. The coefficient of variation of variation for intercepts is reduced, as predicted, from 65.8 to 54.33.

Table 1 Taxonomic and Model II RMA regression statistics for production relationships $D^{2/3}(0)$ versus $D^{2/3}(20)$ for 45 species of tropical trees

Species	Family	<i>n</i>	ρ	Slope	Intercept	r^2	95%
<i>Agonandra macrocarpa</i> L.O. Wms.	Opiliaceae	19	0.74	0.978	1.484	0.874	0.759–1.196
<i>Albizia adinocephala</i> (Donn.Sm.) Britt. & Rose	Fabaceae	13	–	1.059	1.521	0.685	0.664–1.453
<i>Allophylus occidentalis</i> (Sw.) Radlk.	Sapindaceae	7	–	0.987	0.840	0.820	0.505–1.470
<i>Annona reticulata</i> L.	Annonaceae	13	0.57	0.981	1.046	0.826	0.709–1.252
<i>Astronium graveolens</i> Jacq.	Anacardiaceae	73	0.74	0.994	1.127	0.903	0.921–1.067
<i>Ateleia herbert-smithii</i> Pittier	Fabaceae	24	–	1.008	1.854	0.935	0.894–1.121
<i>Bombacopsis quinata</i> (Jacq.) Dugand	Bombaceae	23	0.47	1.000	2.327	0.649	0.731–1.269
<i>Bursera simaruba</i> (L.) Sarg.	Burseraceae	191	0.495	0.935	2.838	0.812	0.877–1.000
<i>Calyophyllum candidissimum</i> (Vahl.) DC.	Rubiaceae	83	0.685	0.926	2.175	0.887	0.857–1.000
<i>Capparis indica</i> (L.) Fawc. & Druce	Capparidaceae	19	–	1.090	0.485	0.815	0.850–1.329
<i>Cecropia peltata</i> L.	Moraceae	12	0.30	1.173	1.307	0.549	0.589–1.756
<i>Cedrela odorata</i> L.	Meliaceae	38	0.49	1.002	2.714	0.883	0.887–1.118
<i>Chomelia spinosa</i> Jacq.	Rubiaceae	65	0.72	1.121	0.174	0.672	0.959–1.282
<i>Cochlospermum vitifolium</i> (Wild.) Spreng.	Cochlospermaceae	90	0.19	1.067	2.607	0.600	0.859–1.280
<i>Cordia alliodora</i> (R. & P.) Oken	Boraginaceae	43	0.651	1.245	0.499	0.273	0.910–1.580
<i>Cordia panamensis</i> Riley	Boraginaceae	32	–	0.978	1.235	0.727	0.787–1.169
<i>Dilodendron costaricense</i> (Radlk.) Gentry & Steyerf.	Sapindaceae	10	–	0.944	2.210	0.887	0.685–1.202
<i>Enterolobium cyclocarpum</i> (Jacq.) Griseb.	Fabaceae	13	0.44	1.131	2.024	0.663	0.695–1.566
<i>Exostema mexicanum</i> A. Gray	Rubiaceae	126	–	0.952	2.248	0.703	0.860–1.044
<i>Guarea glabra</i> Vahl.	Meliaceae	7	–	0.958	0.806	0.980	0.802–1.114
<i>Guazuma ulmifolia</i> Lam.	Sterculiaceae	233	0.635	1.073	1.033	0.768	1.000–1.140
<i>Guettarda macrosperma</i> Don. Sm.	Rubiaceae	43	–	0.882	3.053	0.846	0.772–0.991
<i>Hemilangium excelsum</i> (HBK.) Ac.Smith	Hippocrateaceae	61	–	1.121	0.109	0.713	0.965–1.280
<i>Licania arborea</i> Seem.	Chrysobalanaceae	33	0.675	1.086	0.507	0.966	1.000–1.160
<i>Lonchocarpus costaricensis</i> (Con.Sm.) Pitt	Fabaceae	11	0.62	1.109	0.710	0.858	0.794–1.424
<i>Luehea speciosa</i> Wild.	Tiliaceae	381	0.91	1.100	0.589	0.902	1.060–1.131
<i>Machaerium biovulatum</i> Mecheli	Fabaceae	21	0.90	1.028	0.828	0.826	0.822–1.234
<i>Maclura tinctoria</i> (L.) Don	Moraceae	81	0.485	1.060	1.983	0.753	0.943–1.180
<i>Manilkara chicle</i> (Pittier) Gilly	Sapotaceae	37	0.45	0.979	1.411	0.931	0.890–1.067
<i>Myrospermum frutescens</i> Jacq.	Fabaceae	11	0.705	0.998	1.876	0.903	0.763–1.233
<i>Ocotea veraguensis</i> (Meisn.) Mez	Lauraceae	31	–	1.094	0.171	0.742	0.883–1.300
<i>Pisonia macranthocarpa</i> Donn.Smith	Nyctangaceae	28	–	1.009	0.637	0.857	0.855–1.163
<i>Pithecellobium saman</i> (Jacq.) Benth	Fabaceae	11	0.645	1.056	0.544	0.919	0.829–1.283
<i>Rehdera trinervis</i> (Blake) Mold.	Verbenaceae	17	–	0.850	2.691	0.913	0.713–0.988
<i>Sapium thelocarpum</i> Schm. & Pitt.	Euphorbiaceae	34	0.50	1.296	0.383	0.668	1.022–1.570
<i>Schoepfia schreberi</i> J.F. Gemel.	Olacaceae	17	0.70	1.111	0.485	0.785	0.828–1.395
<i>Sciadodendron excelsum</i> Griseb.	Arillaceae	20	–	1.074	2.882	0.715	0.790–1.359
<i>Sideroxylon capiri</i> (A.D.C.) Pitter	Sapotaceae	10	0.67	1.050	1.142	0.901	0.780–1.320
<i>Simarouba glauca</i> DC.	Simaroubaceae	20	0.68	1.028	1.100	0.932	0.900–1.564
<i>Spondias mombin</i> L.	Anacardiaceae	144	0.395	0.906	2.772	0.723	0.827–0.985
<i>Tabebuia ochracea</i> Standl.	Bignoniaceae	55	0.975	1.050	0.760	0.852	0.935–1.156
<i>Tabebuia rosea</i> (Bertol.) DC.	Bignoniaceae	12	0.75	1.060	0.778	0.953	0.897–1.219
<i>Trichilia americana</i> (Sesse & Mocino) T.D. Penn.	Meliaceae	20	0.44	1.004	1.741	0.778	0.770–1.239
<i>Trichilia cuneata</i> Radlk.	Meliaceae	39	–	0.877	2.240	0.725	0.724–1.030
<i>Zuelania guidonia</i> (Sw.) Britton & Millsp.	Flacortiaceae	12	–	1.292	–0.335	0.860	0.951–1.633

Note that 95% confidence intervals for essentially all species include the predicted slope of 1 indicating that production $\propto M^{3/4}$.

the optimal life history yield the approximate rule

$$E_{\alpha} \approx \frac{(1.3)\rho^{3/4}M_{\alpha}^{1/4}}{C_G C_B} \quad (1)$$

for the average adult lifespan, E_{α} , is related to size at maturation, M_{α} . Equation (1) states that between species with similar M_{α} , E_{α} is proportional to $\rho^{3/4}$. If wood density is constant, this also implies that E_{α} is proportional to $M_{\alpha}^{1/4}$. Although we are not aware of any data sets precise enough to test this prediction, qualitative support is provided by several studies showing a generally positive relationship between wood density, lifespan and age to reproduction^{4,11,19,22–24}. In the future, one could relax the assumption $\lambda = 1$ and explore the general case.

The framework developed above highlights the central roles of allometric scaling and wood density in the life histories of trees. The tropical tree species studied here varied from fast-growing disturbance specialists with low-density wood and short life spans to slow growing emergent trees with dense wood and long life spans. Despite all of this variation, production scaled as $M^{3/4}$, the same as in animals^{1,14}. Although all of the tree species produced new biomass at nearly the same rates (Fig. 2b), differences in wood density resulted in substantial differences in growth rates as measured as change in basal diameter (Fig. 2a). Wood density is frequently cited as being a 'principal determinant' of life history variation in woody plants^{11,19,22–26}. It is correlated with stem water storage and transport capacity, resistance to decay and leaf characteristics such as toughness and deciduousness^{4,11,19,22–26}. Although the size of trees powerfully constrains rates of carbon

fixation, differences in carbon allocation strategies facilitate the coexistence of multiple species in tropical forests owing to variation in growth rate, lifespan and reproductive effort¹⁰.

We have shown that a general allometric framework incorporating the specifics of vascular transport and allocation can account for numerous features of biological diversity^{14–16}. Because of the availability of data to evaluate the model, we have focused on scaling of growth and reproduction in trees; however, the model could easily be modified for plants with different growth forms (herbs) and life histories (monocarpic). Many details of plant anatomy, physiology, and ecology (for example, water and nutrient availability, plant density) can be incorporated into the allometric coefficients (the C s) to link pattern and process across multiple scales of biological organization. Interspecific variation in allocation to roots or defences should also be reflected in the C values. Production is ultimately limited by the physical and biological constraints that limit transportation through the vascular system. However, species differ in how they allocate production. These differences in rates of growth, life span, time until reproduction, wood density and other variables can be mechanistically linked to a general allometric framework of allocation. More importantly, this framework shows how size-dependent variation in life-history strategies can be derived from a set of general allometric principles. □

Methods

Study area

Measurements of diameters at breast height (dbh) of trees were recorded within a permanently marked study plot of seasonally dry tropical forest (10° 45' N, 85° 30' W)

within Sector Santa Rosa, Area de Conservacion, Guanacaste (ACG), of northwest Costa Rica⁷. In 1976, all stems ≥ 3 cm dbh were mapped within a continuous 680 m $\times$ 240 m (16.32 Ha) area of forest²⁰ by S. P. Hubbell. Using an identical mapping protocol, a second remap of the San Emilio forest was completed between 1995 and 1996. In total, 46,833 individuals have been surveyed, 26,960 in 1976 and 19,873 in 1996. Together, the two surveys document 20 yr of growth and population change for about 150 species. The plot is composed of secondary growth forest and is heterogeneous with respect to age, topography and degree of deciduousness.

Calculation of individual tree growth

In 1976, most trees greater than 10 cm dbh were tagged with aluminum tree markers and given a unique identification number. Because few smaller individuals were given aluminum tags in 1976, tree growth was usually followed only for those trees greater than 10 cm dbh. Growth was calculated by monitoring changes in dbh for each individual. To ensure an accurate estimate of growth, a species was included only if a minimum representation of seven individuals had initial stem diameters ≥ 10 cm, and the diameter range of all individuals ≥ 20 cm. As the minimum diameter cut off for individuals was 10 cm, this imposed a minimum size range of 30 cm. Only individuals experiencing positive growth in the 20-year period were used for the calculation of allometric equations. In some cases, individuals experienced no change or even a decrease in diameter over time. This was usually due to partial death, loss of the main trunk or measuring errors. The 45 species meeting the above criteria are listed in Table 1. Production equations for each species were generated by plotting $D^{2/3}(0)$ versus $D^{2/3}(20)$ on linear axes. Because dbh was measured identically in 1976 and 1996, measurement error is likely to be equally distributed across the x and y axes. For these reasons, allometric slopes were determined using Model II RMA regression^{1,28,29}. Equations and statistics for each species are also reported in Table 1.

Species-specific wood density

The specific wood density, ρ , is a simple measure of the total dry mass per unit volume of wood (g cm^{-3}). The specific density of wood is closely related to mechanical properties of strength, such as elastic moduli, which describe resistance to static and impact bending, compression and tension²⁸. For 29 of the 45 species reported in this study, values of specific wood density, ρ , in g cm^{-3} , were taken from the literature^{24,26,30}. If more than one study reported a different value for a species, then the average value was used (Table 1).

Received 9 June; accepted 12 August 1999.

- Charnov, E. L. *Life History Invariants: Some Explorations of Symmetry in Evolutionary Ecology* (Oxford Univ. Press, Oxford, 1993).
- Stearns, S. C. *The Evolution of Life Histories* (Oxford Univ. Press, Oxford, 1992).
- Richards, P. W. *The Tropical Rain Forest* 2nd edn (Cambridge Univ. Press, Cambridge, 1996).
- Chambers, J. Q., Higuchi, N. & Schimel, J. P. Ancient trees in Amazonia. *Nature* **391**, 135–136 (1998).
- Grime, J. P. & Hunt, R. Relative growth-rate: its range and adaptive significance in a local flora. *J. Ecology* **63**, 393–422 (1975).
- Tilman, D. *Plant Strategies and the Dynamics and Structure of Plant Communities* (Princeton Univ. Press, Princeton, 1988).
- Cebrian, J. & Duarte, C. M. The dependence of herbivory on growth rate in natural plant communities. *Funct. Ecol.* **8**, 518–525 (1994).
- Gleason, S. K. & Tilman, D. Plant allocation, growth rate and successional status. *Funct. Ecol.* **8**, 543–550 (1994).
- Ricklefs, R. E. Environmental heterogeneity and plant species diversity: an hypothesis. *Am. Nat.* **111**, 376–381 (1977).
- Grubb, P. J. The maintenance of species diversity in plant communities: the importance of the regeneration niche. *Biol. Rev.* **52**, 107–145 (1977).
- Denslow, J. S. Gap partitioning among tropical rain forest trees. *Biotropica* (Suppl.), **12**, 47–55 (1980).
- Williamson, G. B. Gradients in wood specific gravity of trees. *Bull. Torr. Bot. Club* **111**, 51–55 (1996).
- Hubbell, S. P. *et al.* Light-gap disturbances, recruitment limitation, and tree diversity in a neotropical forest. *Science* **283**, 554–557 (1999).
- West, G. B., Brown, J. H. & Enquist, B. J. A general model for the origin of allometric scaling laws in biology. *Science* **276**, 122–126 (1997).
- Enquist, B. J., Brown, J. H. & West, G. B. Allometric scaling of plant energetics and population density. *Nature* **395**, 163–165 (1998).
- West, G. B., Brown, J. H. & Enquist, B. J. A general model for the structure and allometry of plant vascular systems. *Nature* **400**, 664–667 (1999).
- Peters, R. H. *et al.* The allometry of the weight of fruit on trees and shrubs in Barbados. *Oecologia* **74**, 612–616 (1988).
- Niklas, K. The allometry of plant reproductive biomass and stem diameter. *Am. J. Bot.* **80**, 461–467 (1993).
- Thomas, S. C. Reproductive allometry in Malaysian rain forest trees: biomechanics versus optimal allocation. *Evol. Ecol.* **10**, 517–530 (1996).
- Stevens, G. C. Lianas as structural parasites: the *Bursera simaruba* example. *Ecol.* **68**, 77–81 (1987).
- Whittaker, R. H. & Woodwell, G. M. Dimension and production relations of trees and shrubs in the Brookhaven Forest, New York. *Ecology* **56**, 1–25 (1968).
- Smith, D. W. & Tunney, P. R. Specific density and caloric value of the trunk wood of white birch, black cherry, and sugar maple and their relationship to forest succession. *Can. J. For. Res.* **12**, 186–190 (1982).
- Augsburger, C. K. Seed dispersal of the tropical tree *Platyposidium elegans* and the escape of its seedlings from fungal pathogens. *J. Ecol.* **71**, 759–771 (1983).
- Borchert, R. Soil and stem water storage determine phenology and distribution of Dry Tropical forest trees. *Ecology* **75**, 1437–1449 (1994).
- Sobrado, M. A. Aspects of tissue water relations of evergreen and seasonal changes in leaf water potential components of evergreen and deciduous species coexisting in tropical forests. *Oecologia* **68**, 413–416 (1986).
- Fearnside, P. M. Wood density for estimating forest biomass in Brazilian Amazonia. *For. Ecol. Manage.*

90, 59–87 (1997).

- Janzen, D. H. *Guanacaste National Park: Tropical Education, and Cultural Restoration* (Editorial Univ. Estatal a Distancia, San Jose, 1986).
- Niklas, K. J. *Plant Allometry: The Scaling of Form and Process* (Univ. Chicago Press, Chicago, 1994).
- Harvey, P. H. & Pagel, M. D. *The Comparative Method in Evolutionary Biology* (Oxford Univ. Press, Oxford, 1991).
- Malavassi, I. C. *Maderas de Costa Rica: 150 Especies Forestales* (Univ. de Costa Rica, San Jose, 1998).

Acknowledgements

We thank R. J. Whittaker, G. C. Stevens, D. H. Janzen, J. J. Sullivan, L. Brown, C. A. F. Enquist, A. Masis and the A.C.G. for comments and help with data collection. B.J.E. was supported by a NSF postdoctoral fellowship, G.B.W. by the US Department of Energy and the NSF, E.L.C. by a MacArthur fellowship and J.H.B. by a University of New Mexico Faculty Research Semester. B.J.E., G.B.W. and J.H.B. were also supported by the Thaw Charitable Trust.

Correspondence and requests for materials should be addressed to B.J.E. (e-mail: enquist@nceas.ucsb.edu).

Optimizing the success of random searches

G. M. Viswanathan*†‡, Sergey V. Buldyrev*, Shlomo Havlin*§, M. G. E. da Luz||, E. P. Raposo||# & H. Eugene Stanley*

* Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215, USA

† International Center for Complex Systems and Departamento de Física Teórica e Experimental, Universidade Federal do Rio Grande do Norte, 59072-970, Natal-RN, Brazil

‡ Departamento de Física, Universidade Federal de Alagoas, 57072-970, Maceió-AL, Brazil

§ Gonda-Goldschmied Center and Department of Physics, Bar Ilan University, Ramat Gan, Israel

|| Lyman Laboratory of Physics, Harvard University, Cambridge, Massachusetts 02138, USA

¶ Departamento de Física, Universidade Federal do Paraná, 81531-970, Curitiba-PR, Brazil

Laboratório de Física Teórica e Computacional, Departamento de Física, Universidade Federal de Pernambuco, 50670-901, Recife-PE, Brazil

We address the general question of what is the best statistical strategy to adapt in order to search efficiently for randomly located objects ('target sites'). It is often assumed in foraging theory that the flight lengths of a forager have a characteristic scale: from this assumption gaussian, Rayleigh and other classical distributions with well-defined variances have arisen. However, such theories cannot explain the long-tailed power-law distributions^{1,2} of flight lengths or flight times^{3–6} that are observed experimentally. Here we study how the search efficiency depends on the probability distribution of flight lengths taken by a forager that can detect target sites only in its limited vicinity. We show that, when the target sites are sparse and can be visited any number of times, an inverse square power-law distribution of flight lengths, corresponding to Lévy flight motion, is an optimal strategy. We test the theory by analysing experimental foraging data on selected insect, mammal and bird species, and find that they are consistent with the predicted inverse square power-law distributions.

Lévy flights are characterized by a distribution function

$$P(l_j) \sim l_j^{-\mu} \quad (1)$$

with $1 < \mu \leq 3$, where l_j is the flight length. The gaussian is the stable distribution for the special case $\mu \geq 3$ owing to the central-limit theorem, while values $\mu \leq 1$ do not correspond to probability distributions that can be normalized². This generalization, equation (1), introduces a natural parameter μ such that we essentially have a

family of distributions. Our strategy is to find the value of the parameter—and hence the distribution—that optimizes the search process. Levandowsky *et al.*^{3,4} have suggested why microorganisms may perform Lévy flights. A Lévy distribution is advantageous when target sites are sparsely and randomly distributed, irrespective of the value of μ chosen⁷, because the probability of returning to a previously visited site is smaller than for a gaussian distribution. Another explanation, proposed by Shlesinger⁶, argues that foragers may perform Lévy flights because the number of new visited sites is much larger for N Lévy walkers than for N brownian walkers^{8–11}. A Lévy flight strategy is also a good solution for the related problem of where to locate N radar stations to optimize the search for M targets¹².

Here we develop an idealized model which captures some of the essential dynamics of foraging in the limiting case in which predator–prey relationships are ignored and learning is minimized. We assume that target sites are distributed randomly, and that the forager behaves as follows (see Fig. 1):

(1) If a target site lies within a ‘direct vision’ distance r_v , then the forager moves on a straight line to the nearest target site. A finite value of r_v , no matter how large, models the constraint that no forager can detect (or ‘remember’) a target site located an arbitrarily large distance away.

(2) If there is no target site within a distance r_v , then the forager chooses a direction at random and a distance l_j from the probability distribution (equation (1)). It then incrementally moves to the new point, constantly looking for a target within a radius r_v along its way. If it does not detect a target, it stops after traversing the distance l_j and chooses a new direction and a new distance l_{j+1} ; otherwise, it proceeds to the target as in rule (1).

In the case of non-destructive foraging, the forager can visit the same target site many times. Non-destructive foraging can occur in either of two cases: if the target sites become temporarily depleted or fall below some fixed concentration threshold, and if the forager becomes satiated and leaves the area. In the case of destructive foraging, the target site found by the forager becomes undetectable in subsequent flights.

First, we solve this model analytically. Let λ be the mean free path of the forager between successive target sites (for two dimensions, $\lambda \equiv (2r_v\rho)^{-1}$ where ρ is the target-site area density). The mean flight distance is

$$\langle l \rangle \approx \frac{\int_{r_v}^{\lambda} l^{1-\mu} dl + \lambda \int_{\lambda}^{\infty} l^{-\mu} dl}{\int_{r_v}^{\infty} l^{-\mu} dl} \quad (2)$$

$$= \left(\frac{\mu-1}{2-\mu} \right) \left(\frac{\lambda^{2-\mu} - r_v^{2-\mu}}{r_v^{1-\mu}} \right) + \frac{\lambda^{2-\mu}}{r_v^{1-\mu}}$$

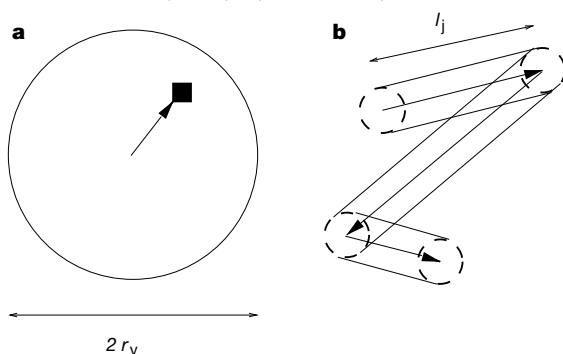


Figure 1 Foraging strategy. **a**, If a target site (solid square) is located within a ‘direct-vision’ distance r_v , then the forager moves on a straight line to it. **b**, If there is no target site within a distance r_v , then the forager chooses a random direction and a random distance l_j from the Lévy probability distribution $P(l_j) \sim l_j^{-\mu}$, and then proceeds as described in the text.

The second term of this ‘mean field’ calculation is approximate because it assumes that the distances λ_k between successive sites k are all equal to λ . The probability distribution has a finite cutoff λ and corresponds to a truncated Lévy distribution. An infinite λ leads to divergences for $\mu \leq 2$ (see Fig. 2a). The cutoff causes convergence to gaussian behaviour only after a very large number of steps¹³. A more rigorous treatment that considers a Poisson distribution of λ_k does not alter the results significantly (see simulation results below).

We define the search efficiency function $\eta(\mu)$ to be the ratio of the number of target sites visited to the total distance traversed by the forager, so that

$$\eta = \frac{1}{\langle l \rangle N} \quad (3)$$

where N is the mean number of flights taken by a Lévy forager while

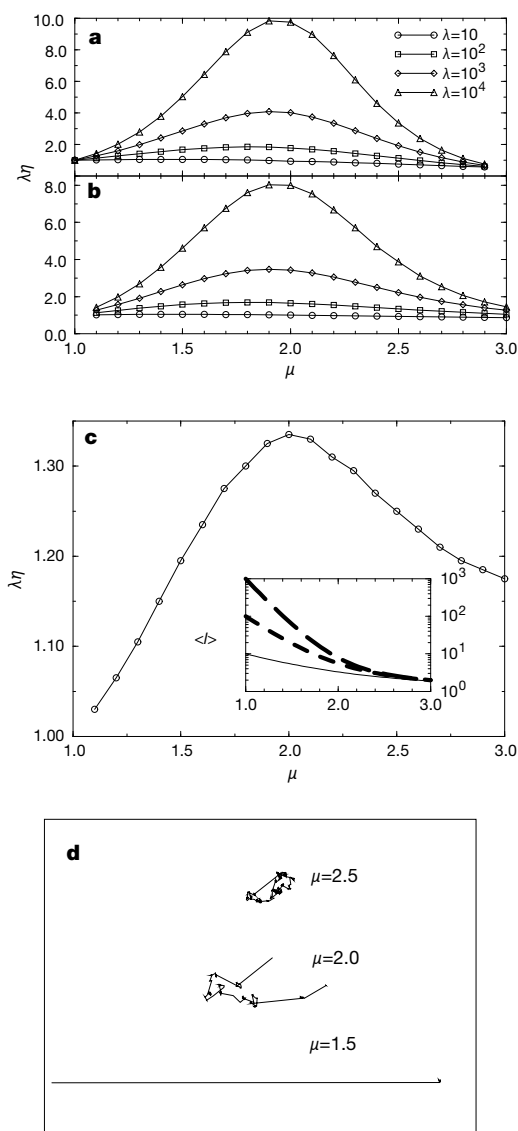


Figure 2 a, b, The product of the mean free path λ and the foraging efficiency η against the Lévy parameter μ in one dimension for different values of λ , found from equations (2) and (3) ($r_v = 1$) for the case of non-destructive foraging (**a**) and from simulations (**b**). **c**, $\lambda\eta$ found from simulations in two dimensions, with $\lambda = 5,000$ ($r_v = 1$). In each case, $\mu_{opt} \approx 2$ emerges as an optimal value of the Lévy flight exponent. Inset shows $\langle l \rangle$ as a function of μ for $r_v = 1$ and $\lambda = 10$ (solid line), $\lambda = 10^2$ (dashed), $\lambda = 10^3$ (long-dashed). The results indicate that flights become too long when $\mu < 2$, causing inefficient foraging (see equation (3)). **d**, Two-dimensional random walks for $\mu = 2.5$, 2.0 and 1.5 with identical total lengths of 10^3 units.

travelling between two successive target sites. A low value of η can result from either a larger N or a large $\langle l \rangle$, corresponding to large and small μ , respectively. For intermediate values of μ it is thus conceivable that a maximum in η can arise. We first consider the case of destructive foraging. The mean number of flights N_d taken to travel an average distance λ between two successive target sites scales¹⁴ as

$$N_d \approx (\lambda/r_v)^{\mu-1} \quad (4)$$

for $1 < \mu \leq 3$. Here $\mu - 1$ is the fractal dimension of the set of sites

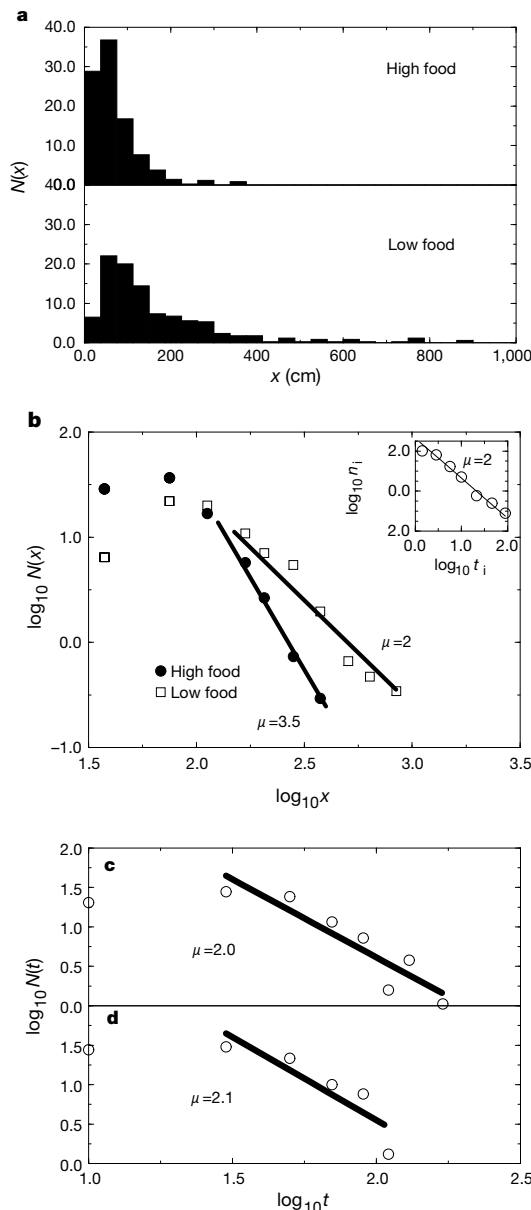


Figure 3 Foraging by bumble-bees and deer. **a**, Flight-length percentage distributions for foraging bumble-bees. We digitized the data from ref. 15. **b**, Double-log plot of the same data; the value $\mu \approx 2$ for low nectar concentration is the same as predicted by the model. We are interested solely in the long flights, because the power-law exponent μ is not affected by short flights. The value $\mu \approx 3.5$ for higher nectar concentrations (approximately 10 times) in which long flights become very rare (see text) is consistent with the prediction that η becomes independent of μ when $\lambda \leq r_v$. We smoothed the data using running averaging. The inset displays a double-log plot of the histograms of flight times (in 1-h intervals) for the wandering albatross⁶. **c**, **d**, Double-log plot of the foraging time (in s) percentage distributions for deer in wild areas (**c**) and fenced areas (**d**). We digitized the original data from ref. 16. In fenced areas, spatial limitation introduces an artificially larger number of 'turnings'.

visited by a Lévy random walker. (If the number of sites m in a closed region of a radius r scales as $m \approx r^{d_f}$, then d_f is the fractal dimension of the set of sites.) Note that $N_d \approx (\lambda/r_v)^2$ for $\mu \geq 3$ (brownian motion)². We also consider the case of non-destructive foraging. Because previously visited sites can be revisited, equation (4) overestimates the mean number N_n of flights between successive target sites for the non-destructive case. We show below that $N_n \approx N_d^{1/2}$. Let r_o be the small distance between the last visited target site and the position after the first subsequent flight. For a brownian walker, in the case of destructive foraging, $N_d \approx \lambda^2$ because the average time required for a random walker in one dimension who is initially in the middle of a container of radius λ to reach the boundary is $N_d = \lambda^2/(2D)$, where D is the diffusion constant. However, for the non-destructive case, $N_n = (\lambda - r_o)r_o/(2D)$, because the previous site (only a small distance r_o away) can be revisited—that is, the scaling is quadratic in the former case and linear in the latter. We have found this result also to hold for anomalous diffusion and spatial dimensions higher than 1. It follows that

$$N_n \approx (\lambda/r_v)^{(\mu-1)/2} \quad (5)$$

for $1 < \mu \leq 3$. We have also systematically tested equation (5) using simulations, and find that the approximation becomes increasingly better as (λ/r_v) increases (compare also Fig. 2a, b).

Having found expressions for N_d and N_n , we first consider the case in which the target sites are plentiful, that is, $\lambda \leq r_v$. Then $\langle l \rangle \approx \lambda$ and $N_d \approx N_n \approx 1$. Hence, η becomes independent of μ . This behaviour does not correspond to Lévy flight motion because long flights with $l_j \gg r_v$ are practically non-existent. We next study the more usual case in which the target sites are sparsely distributed, defined by $\lambda \gg r_v$. Substituting equations (2) and (4) into equation (3), we find for destructive foraging that the mean efficiency η has no maximum, with lower values of μ leading to more efficient foraging. Note that when $\mu = 1 + \epsilon$ with $\epsilon \rightarrow 0^+$, the fraction of flights with $l_j < \lambda$ becomes negligible, and effectively the forager moves along straight lines until it detects a target site. For non-destructive foraging, we note that if $\lambda \gg r_v$, then $N_d \gg N_n$. Substituting equations (2) and (5) into equation (3), and differentiating with respect to μ , we find that the efficiency $\eta = 1/(N_n \langle l \rangle)$ is optimum at

$$\mu_{\text{opt}} = 2 - \delta \quad (6)$$

where $\delta \approx 1/[\ln(\lambda/r_v)]^2$. So, in the absence of *a priori* knowledge about the distribution of target sites, an optimal strategy for non-destructive foraging is to choose $\mu_{\text{opt}} = 2$ when λ/r_v is large but not exactly known.

We test the above theoretical results with numerical simulations, which have the advantage that no approximations are made. Specifically, we perform one- and two-dimensional simulations of the model and study how η varies with μ for the case of non-destructive foraging for a random distribution of target sites. Figure 2a shows that the simulation agrees with the analytical results (Fig. 2b), and approaches $\mu_{\text{opt}} = 2$ as $\lambda \rightarrow \infty$. The discrepancy in η near $\mu = 3$ is due to the slow convergence of N_n , which approaches the expected scaling behaviour as $\lambda \rightarrow \infty$. The simulation results for two dimensional non-destructive foraging also show maxima near $\mu_{\text{opt}} = 2$. Figure 2c shows simulated foraging in a system of size $10^4 \times 10^4$ with $r_v = 1$, periodic boundary conditions and $\lambda/r_v = 5 \times 10^3$. For destructive foraging with $\lambda \gg r_v$, simulations show that $\mu \rightarrow 1$ optimizes the efficiency as predicted. In contrast, if the target sites are densely distributed such that $\lambda \approx r_v$, then, as expected, we find no significant effect of varying μ . Our findings agree with the theoretical predictions and raise the possibility that Lévy-flight foraging with $\mu < 3$ may be confined to instances of low global target-site concentration, as the advantage of long flights becomes negligible when there are ample target sites (see also Figs 2b and 3b). Note that our simulation results do not suffer from the approximations inherent in equation (2).

We compare our analytical and simulation results with foraging

data on a variety of animals. The original foraging data on bees (Fig. 3a) were collected by recording the landing sites of individual bees¹⁵. We find that, when the nectar concentration is low, the flight-length distribution decays as in equation (1), with $\mu \approx 2$ (Fig. 3b). (The exponent μ is not affected by short flights.) We also find the value $\mu \approx 2$ for the foraging-time distribution of the wandering albatross⁶ (Fig. 3b, inset) and deer (Fig. 3c, d) in both wild and fenced areas¹⁶ (foraging times and lengths are assumed to be proportional). The value $2 \leq \mu \leq 2.5$ found for amoebas⁴ is also consistent with the predicted Lévy-flight motion.

The above theoretical arguments and numerical simulations suggest that $\mu \approx 2$ is the optimal value for a search in any dimension. This is analogous to the behaviour of random walks whose mean-square displacement is proportional to the number of steps in any dimension¹⁷. Furthermore, equations (4) and (5) describe the correct scaling properties even in the presence of short-range correlations in the directions and lengths of the flights. Short-range correlations can alter the width of the distribution $P(l)$, but cannot change μ , so our findings remain unchanged. Hence, learning, predator-prey relationships and other short-term memory effects become unimportant in the long-time, long-distance limit. A finite λ ensures that the longest flights are not energetically impossible. Our findings may also be relevant to the study of population dynamics. Specifically, each value of μ is related to a different type of redistribution kernel¹⁸; for example, $\mu \geq 3$ corresponds to the normal (or similar) distribution, while $\mu = 2$ corresponds to a Cauchy distribution (see also ref. 19). Finally, note that non-destructive foraging is more realistic than destructive foraging because, in nature, 'targets' such as flowers, fish and berries are often found in patches that regenerate. Organisms are often in clusters for reproductive purposes, and sometimes such clusters may have fractal shapes²⁰. Thus, the forager can revisit the same food patch many times. We simulated destructive foraging in various patchy and fractal target-site distributions and found results consistent with non-destructive foraging with uniformly distributed target sites. □

Received 10 May; accepted 12 August 1999.

1. Tsallis, C. Lévy distributions. *Phys. World* **10**, 42–45 (1997).
2. Schlesinger, M. F., Zaslavsky, G. M. & Frisch, U. (eds) *Lévy Flights and Related Topics in Physics* (Springer, Berlin, 1995).
3. Levandowsky, M., Klafter, J. & White, B. S. Swimming behavior and chemosensory responses in the tropical microzooplankton as a function of the hydrodynamic regime. *Bull. Mar. Sci.* **43**, 758–763 (1988).
4. Schuster, F. L. & Levandowsky, M. Chemosensory responses of *Acanthamoeba castellanii*: Visual analysis of random movement and responses to chemical signals. *J. Eukaryotic Microbiol.* **43**, 150–158 (1996).
5. Cole, B. J. Fractal time in animal behaviour: The movement activity of *Drosophila*. *Anim. Behav.* **50**, 1317–1324 (1995).
6. Viswanathan, G. M. et al. Lévy flight search patterns of wandering albatrosses. *Nature* **381**, 413–415 (1996).
7. Schlesinger, M. F. & Klafter, J. in *On Growth and Form* (eds Stanley, H. E. & Ostrowsky, N.) 279–283 (Nijhoff, Dordrecht, 1986).
8. Berkolaiko, G., Havlin, S., Larralde, H. & Weiss, G. H. Expected number of distinct sites visited by N discrete Lévy flights on a one-dimensional lattice. *Phys. Rev. E* **53**, 5774–5778 (1996).
9. Berkolaiko, G. & Havlin, S. Territory covered by N Lévy flights on d -dimensional lattices. *Phys. Rev. E* **55**, 1395–1400 (1997).
10. Larralde, H., Trunfio, P., Havlin, S., Stanley, H. E. & Weiss, G. H. Territory covered by N diffusing particles. *Nature* **355**, 423–426 (1992).
11. Larralde, H., Trunfio, P., Havlin, S., Stanley, H. E. & Weiss, G. H. Number of distinct sites visited by N random walkers. *Phys. Rev. A* **4**, 7128–7138 (1992).
12. Szu, H. in *Dynamic Patterns in Complex Systems* (eds Kelso, J. A. S., Mandell, A. J. & Shlesinger, M. F.) 121–136 (World Scientific, Singapore, 1988).
13. Mantegna, R. N. & Stanley, H. E. Stochastic process with ultra-slow convergence to a gaussian: The truncated Lévy flight. *Phys. Rev. Lett.* **73**, 2946–2949 (1994).
14. Shlesinger, M. F. & Klafter, J. Comment on "Accelerated diffusion in Josephson junctions and related chaotic systems". *Phys. Rev. Lett.* **54**, 2551 (1985).
15. Heinrich, B. Resource heterogeneity and patterns of movement in foraging bumble-bees. *Oecologia* **40**, 235–245 (1979).
16. Focardi, S., Marcellini, P. & Montanaro, P. Do ungulates exhibit a food density threshold—a field-study of optimal foraging and movement patterns. *J. Anim. Ecol.* **65**, 606–620 (1996).
17. Berg, H. C. *Random Walks in Biology* (Princeton Univ. Press, Princeton, 1983).
18. Kot, M., Lewis, M. & van der Driftsch, P. Dispersal data and the spread of invading organisms. *Ecology* **77**, 2027–2042 (1996).
19. Schulman, L. S. *Time's Arrows and Quantum Measurement* (Cambridge Univ. Press, Cambridge, 1997).
20. Sugihara, G. & May, R. Applications of fractals in ecology. *Trends Ecol. Evol.* **5**, 79–86 (1990).

Acknowledgements

We thank V. Afanasyev, N. Dokholyan, I. P. Fittipaldi, P. Ch. Ivanov, U. Laino, L. S. Lucena, E. G. Murphy, P. A. Prince, M. F. Shlesinger, B. D. Stosic and P. Trunfio for discussions, and CNPq, NSF and NIH for financial support.

Correspondence should be addressed to G.M.V. (e-mail: gandhi@fis.ufal.br).

Water stress inhibits plant photosynthesis by decreasing coupling factor and ATP

W. Tezara*, V. J. Mitchell†, S. D. Driscoll† & D. W. Lawlor†

* Instituto de Biología Experimental, Facultad de Ciencias, Universidad Central de Venezuela, Calla Suapora, Colinas de Bello Monte, Apartado 47114, Caracas 1041a, Venezuela

† Biochemistry and Physiology Department, IACR-Rothamsted, Harpenden, Hertfordshire AL5 2JQ, UK

Water stress substantially alters plant metabolism, decreasing plant growth and photosynthesis^{1–4} and profoundly affecting ecosystems and agriculture, and thus human societies⁵. There is controversy over the mechanisms by which stress decreases photosynthetic assimilation of CO₂. Two principal effects are invoked^{2,4}: restricted diffusion of CO₂ into the leaf, caused by stomatal closure^{6–8}, and inhibition of CO₂ metabolism^{9–11}. Here we show, in leaves of sunflower (*Helianthus annuus* L.), that stress decreases CO₂ assimilation more than it slows O₂ evolution, and that the effects are not reversed by high concentrations of CO₂^{12,13}. Stress decreases the amounts of ATP^{9,11} and ribulose biphosphate found in the leaves, correlating with reduced CO₂ assimilation¹¹, but the amount and activity of ribulose biphosphate carboxylase-oxygenase (Rubisco) do not correlate. We show that ATP-synthase (coupling factor) decreases with stress and conclude that photosynthetic assimilation of CO₂ by stressed leaves is not limited by CO₂ diffusion but by inhibition of ribulose biphosphate synthesis, related to lower ATP content resulting from loss of ATP synthase.

When land plants absorb less water from the environment through their roots than is transpired (evaporated) from their leaves, water stress develops. The relative water content (RWC), water potential (ψ) and turgor of cells are decreased and the concentrations of ions and other solutes in the cells are increased, thereby decreasing the osmotic potential^{1–4}. Stomatal pores in the leaf surface progressively close^{2,6,13,14}, decreasing the conductance to water vapour (g_{H₂O}) and thus slowing transpiration and the rate at which water deficits develop^{1–4,6,11,14}. Also, photosynthetic assimilation of CO₂ (A) decreases, often concomitant with, and frequently ascribed to, decreasing conductance to CO₂ (g_{CO₂}) (refs 2–4, 6, 8). However, decreased A is also considered to be caused by inhibition of the photosynthetic carbon reduction (Calvin) cycle^{1,2,9,11}, although there is uncertainty over which biochemical processes are most sensitive to stress^{2,3,6,15,16}. We assessed whether A is controlled by g_{CO₂} or by metabolic factors by measuring CO₂ and O₂ exchange, using large CO₂ concentrations (up to 0.1 mol mol^{–1}, equivalent to 10% volume/volume) to overcome small g_{CO₂} (refs 8, 12–14), and by determining important indicators of photosynthetic biochemistry (for example, ribulose biphosphate (RuBP) and Rubisco of the Calvin cycle)^{9,10,16,17}. We considered the role of ATP in particular^{2,9,10} because, although inhibition of photophosphorylation has been demonstrated^{9,10} but is not widely accepted^{13,4,6,18,19}, it may explain the decrease in RuBP and A (refs 2, 11).

In the chloroplasts of leaf cells, capture of photons causes electron transport in the thylakoid membranes, evolution of O₂ (refs 2–4)

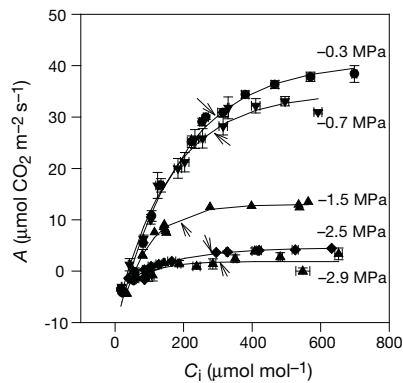


Figure 1 Rate of photosynthetic CO_2 assimilation (A) in sunflower (*H. annuus* L.) as a function of the calculated CO_2 concentration inside the leaf mesophyll (c_i) for leaves of different water potential (ψ , shown against each curve) measured at $800 \mu\text{mol m}^{-2} \text{s}^{-1}$ PAR and 25°C . The c_i values calculated from the particular photosynthetic rate and stomatal conductance (the operating points) at $360 \mu\text{mol CO}_2 \text{mol}^{-1}$ for each water potential are indicated by arrows on the curves.

and generation of NADPH and the transthylakoid proton gradient ($\Delta p\text{H}$) which is responsible for ATP synthesis by ATP synthase (coupling factor, $\text{CF}_1\text{--CF}_0$)^{10,19,20}. The Calvin cycle uses ATP and NADPH to synthesize RuBP, which reacts with CO_2 in a reaction catalysed by Rubisco. Owing to g_{CO_2} , this uptake of CO_2 decreases its concentration within the leaf (c_i) relative to that in the atmosphere (c_a)^{2-4,16,21}: $c_i = c_a - A/g_{\text{CO}_2}$ where $g_{\text{CO}_2} = g_{\text{H}_2\text{O}}/1.6$. The factor 1.6 is the ratio of the diffusivities D of H_2O and CO_2 in air ($D_{\text{H}_2\text{O}}/D_{\text{CO}_2} = 1.6$)^{7,13}. As g_{CO_2} decreases with stress, A may diminish because of small c_i . We tested this by measuring the relationship between A and c_i using large c_a to overcome limiting g_{CO_2} , and the effects of stress on electron transport, by measuring O_2 evolution and NADPH content^{4,18,22,23}. If stress inhibits the Calvin cycle, for example by decreasing Rubisco amounts and activities, then RuBP content will decrease but ATP will increase. If the capacity for ATP synthesis is limiting, indicated by, for example, lower ATP synthase content, then both will decrease^{2,16,18,23}.

The A/c_i curves (Fig. 1) show that both maximum rate of CO_2 assimilation (A_{max}) and carboxylation efficiency, ϕ ($\phi = A/\mu\text{mol CO}_2 \text{mol}^{-1}$), decreased with stress, as observed in wheat¹² and sunflower¹¹. The operating c_i (Fig. 1) decreased as ψ decreased to about -1.5 MPa, but increased with further stress. This shows some stomatal limitation but greatly impaired metabolism. However, if c_i is incorrectly determined this interpretation is invalid^{2-4,14,21,24-26}. Calculation of c_i assumes a uniform and unimodal frequency distribution of g_{CO_2} over the leaf, with no 'patchy stomata'^{14,24-26}. We established that there was no heterogeneity of stomatal conductance by exposing the leaf areas on which A was measured to $^{14}\text{CO}_2$, as previously shown for sunflower¹¹ and in other analyses¹⁴. Calculation of c_i assumes that movement of CO_2 and H_2O across the leaf cuticle, when the stomata are closed, conforms to $D_{\text{H}_2\text{O}}/D_{\text{CO}_2} = 1.6$. However, CO_2 movement may be much restricted²¹. Hence, our calculations of c_i may be erroneous. We recalculated c_i for the most severely stressed leaves using the values of A but assuming different D_{CO_2} values. A 30% decrease in D_{CO_2} progressively reduced c_i , increasing ϕ but not A_{max} . Smaller values of D_{CO_2} led to greatly underestimated c_i and unrealistically increased ϕ , showing that the error is small and does not substantially alter the A/c_i relationships in Fig. 1. This indicates that, despite the very small g_{CO_2} , c_i is not substantially depleted^{2,11} and that photosynthetic metabolism is inhibited by water stress.

A may also be limited by the small conductance of the pathway of CO_2 movement between the intercellular spaces and active sites of Rubisco^{2,3}. To eliminate these, we measured A with increasing CO_2 concentration^{2,4,12,13}, but even at $c_a = 9 \text{ mmol CO}_2 \text{mol}^{-1}$ (0.9%) in a normal gas exchange system we observed no stimulation under

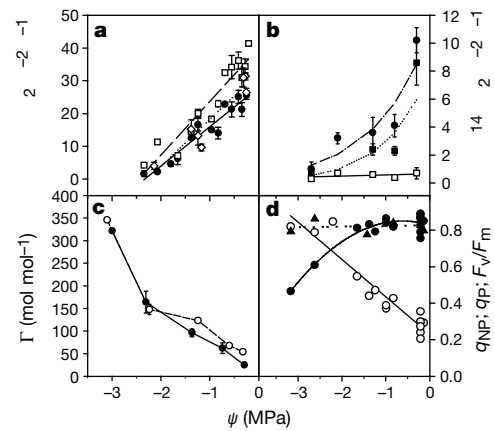


Figure 2 Photosynthetic processes of sunflower (*H. annuus* L.) leaves as a function of leaf water potential (ψ). **a**, Rate of photosynthesis (A) measured at three CO_2 concentrations in the atmosphere: $0.36 \text{ mmol mol}^{-1}$ (0.036%) (circles); 1 mmol mol^{-1} (0.1%) (diamonds); 9 mmol mol^{-1} (0.9%) (squares). **b**, A determined from $^{14}\text{CO}_2$ supplied to leaf discs under conditions used in the oxygen electrode, at CO_2 concentrations of 10 mmol mol^{-1} (1%) (circles); 50 mmol mol^{-1} (5%) (filled squares); $100 \text{ mmol mol}^{-1}$ (10%) (open squares). **c**, Equilibrium CO_2 concentration, Γ , measured at $800 \mu\text{mol m}^{-2} \text{s}^{-1}$ PAR and 25°C in $0.21 \text{ mol mol}^{-1} \text{O}_2$ (21%) (open circles) or $0.02 \text{ mol mol}^{-1} \text{O}_2$ (2%) (filled circles). **d**, Photochemical quenching, q_p (filled circles); non-photochemical quenching, q_{NP} (open circles); F_v/F_m (triangles), derived from chlorophyll a fluorescence measurements.

stress (Fig. 2a). Indeed, A ceased at the same RWC (50%: $\psi \approx -2$ MPa) for all values of c_a . This agrees with the findings of ref. 12 for water-stressed (but not abscisic acid-treated) leaves, but disagrees with those of ref. 8.

Oxygen evolution is considered to be evidence for continued CO_2 metabolism under stress^{3,4,8}. We measured O_2 evolution in different c_a . At $\psi = -0.6$ MPa in $10 \text{ mmol CO}_2 \text{mol}^{-1}$, the rate of O_2 evolution was $29.9 \mu\text{mol m}^{-2} \text{s}^{-1}$, decreasing by $\sim 45\%$ and 60% at $\psi = -1.5$ and -2.5 MPa, respectively. Oxygen evolution was not stimulated by $50 \text{ mmol CO}_2 \text{mol}^{-1}$ (5%) compared with $10 \text{ mmol CO}_2 \text{mol}^{-1}$ at $\psi = -0.6$ MPa, but increased by 28% and 19% at $\psi = -1.5$ and -2.5 MPa, respectively, showing some diffusion limitation. Increasing c_a to $100 \text{ mmol CO}_2 \text{mol}^{-1}$ (10%) completely inhibited O_2 evolution irrespective of stress. We tested whether CO_2 assimilation (A) continues during O_2 evolution with decreasing ψ by measuring $^{14}\text{CO}_2$ assimilation as ^{14}C accumulation in organic compounds at high c_a . A decreased (Fig. 2b) with stress irrespective of CO_2 , and 0.1 mol mol^{-1} (10%) CO_2 completely inhibited CO_2 assimilation as well as O_2 evolution, similar to the results of ref. 11 but contrary to those of ref. 8. If O_2 evolution depends on CO_2 influx and assimilation in the leaf^{8,4}, ^{14}C should have accumulated. However, it did not, despite similar characteristics and pathways of movement of O_2 and CO_2 , indicating that O_2 evolution under stress may not be dependent on Calvin cycle function.

Metabolic inhibition of A is also shown by the progressive increase in the equilibrium CO_2 compensation point (Γ), measured on whole, detached pre-stressed leaves, as stress increased (Fig. 2c). This increase occurred at an O_2 concentration of $0.21 \text{ mol mol}^{-1}$ (that is, current atmospheric concentration of 21%), as observed for *Triticum aestivum* and *H. annuus*^{2,13}. An O_2 concentration ($0.02 \text{ mol mol}^{-1}$ or 2%) that prevents glycolate pathway photorespiration^{2,27} decreased Γ compared with $0.21 \text{ mol O}_2 \text{mol}^{-1}$ over the range $\psi = -0.3$ to -1.8 MPa, but not at smaller (more negative) ψ , and did not prevent Γ from increasing with stress. This may be related to respiration (by the tricarboxylic acid cycle), which, in darkness at $0.21 \text{ mol O}_2 \text{mol}^{-1}$, decreased by about 60% as ψ fell from -0.3 to -3.0 MPa (data not shown), continuing even when A had ceased. This confirms that A decreases more than respiration with increasing stress^{2,9}. Thus, under stress, dark respiration determined Γ (which was also increased by photorespiration) but

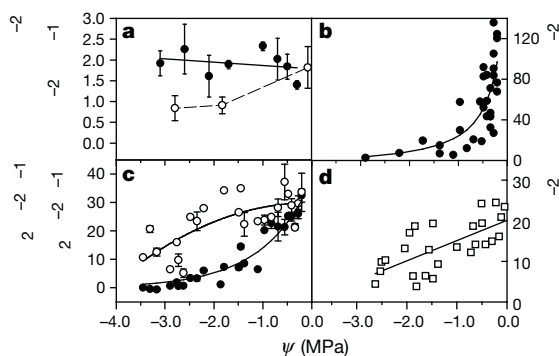


Figure 3 Biochemical composition of sunflower (*H. annuus* L.) leaves as a function of water potential (ψ). **a**, Amounts (m^{-2}) of ribulose biphosphate carboxylase-oxygenase (filled circles) and ATP synthase (open circles). **b**, Amount of ribulose biphosphate. **c**, Rate of photosynthetic CO_2 assimilation, A (filled circles) and the initial activity (init. act.) of ribulose biphosphate carboxylase-oxygenase (open circles). **d**, Amount of ATP.

photorespiration ceased at severe stress. Further evidence for a large change in photosynthetic metabolism caused by stress is provided by chlorophyll *a* fluorescence²⁸, with greatly increasing photochemical quenching (q_{NP}) and decreasing chemical quenching (q_{P}) (refs 2, 4) but no change in the variable-to-maximum fluorescence ratio (F_v/F_m) (Fig. 2d), showing that the efficiency of open photosystem II reaction centres was not impaired. This suggests that photoinhibition^{28,29}, often considered to be a primary effect of stress^{3,4,28} caused by CO_2 limitation, was not important in our experiments, as seen in sunflower²⁹ but not in other studies⁴.

We conclude that water stress does not inhibit A by decreasing CO_2 diffusion (occurring even in elevated CO_2) or by preventing O_2 production and electron transport, but by inhibition of the Calvin cycle.

What in the Calvin cycle is impaired by water stress? Amounts of Rubisco^{2,11,15} per unit leaf area remained approximately constant (Fig. 3a), despite leaf shrinkage, whereas RuBP content (Fig. 3b) was severely decreased by very mild stress, confirming earlier results¹¹. Rubisco activity was decreased only by relatively severe stress (Fig. 3c) and did correlate with the decrease in A (Fig. 3c), although the decrease in A did correlate with the decrease in RuBP as shown¹¹. If Rubisco were inhibited, then RuBP would accumulate relative to the decrease in A , even though the Calvin cycle is autocatalytic and decreased A would decrease synthesis of RuBP^{2,16}. The 3-phosphoglyceric acid (3-PGA) content per unit leaf area was unaffected by water deficit (data not shown). However, this is expected, as respiration, which we have confirmed is relatively less affected by stress than A , produces 3-PGA, thus obscuring Calvin cycle function¹⁶.

ATP content decreased with increasing stress (Fig. 3d), indicating that stress inhibits ATP synthesis rather than the Calvin cycle *per se*^{2,9,10} (disagreeing with ref. 18). If the Calvin cycle is inhibited then ATP content should rise, as in transgenic plants with much decreased phosphoribulokinase activity²³. Decreased ATP indicates that ATP synthesis may be impaired, although increased ATP consumption cannot be excluded². As respiration continues in stressed leaves, although at a smaller rate than in unstressed leaves, it could possibly maintain a relatively large ATP content. However, if the capacity for phosphorylation by dark respiration in leaves is smaller than that for photophosphorylation, decreased ATP is more likely to be caused by inhibition of photophosphorylation^{9,10}. We measured ATP synthase to assess whether loss of this enzyme was related to decreased ATP content. The amount of ATP synthase per m^2 decreased progressively with stress (Fig. 3a). The ATP content per m^2 correlates strongly with the ratio of ATP synthase to Rubisco ($\text{ATP} = -4.07 + 1.91 (\% \text{ ATP synthase/Rubisco})$; $r^2 = 0.99$). We conclude that decreased ATP content is related to a decrease in ATP synthase with increasing stress. Decreasing ATP synthase in leaves by transformation with antisense

had very similar effects to water stress, with increased non-photochemical and decreased photochemical quenching, explained by decreased ATP synthesis¹⁷.

ATP synthase is composed of CF_0 , embedded in the thylakoid membrane, attached to CF_1 , which projects into the chloroplast stroma²⁰. ATP is synthesized by physical rotation of components that alters the conformation of the active enzyme sites, driven by movement of protons through the complex as a consequence of the ΔpH between thylakoid lumen and stroma^{24,20}. Catalytic activity of ATP synthase must depend on very stringent spatial arrangements within and between the subunits. Ionic concentration affects ATP synthase activity and binding between CF_1 and CF_0 *in vivo*²⁰. Large magnesium concentrations decrease photophosphorylation of chloroplasts isolated from stressed leaves^{9,10}, providing a mechanism for the effects of decreasing ψ and RWC on RuBP content and CO_2 assimilation¹¹. We calculate that magnesium concentrations in the chloroplasts of leaves used in our experiment increased twofold between -0.3 and -3 MPa (95% to 50% RWC). As it is the amount of CF_1 that decreases, binding between CF_1 and CF_0 is probably affected, although concentrated Mg^{2+} may also impair catalysis: these aspects require further examination. Differences between ATP synthase from chloroplasts and that from mitochondria, and in their environments²⁰, may underlie the different responses of photosynthesis and respiration to stress. ATP synthesis depends on ΔpH , which is large under stress, judging from the greatly increased q_{NP} (refs 2, 20, 28), and also in leaves with decreased ATP synthase content¹⁷. Loss of CF_1 must block proton flux out of the thylakoid, causing a large ΔpH (and large non-photochemical quenching), so lack of energy does not decrease ATP.

The decrease in A correlates with decreased ATP and ATP synthase. However, ATP content with severe stress was still 30% of that without stress, so A might be expected to continue. This ATP may not be in the chloroplast, or the concentration may be insufficient for RuBP synthesis^{2,11,16,30}. A RuBP concentration below $\sim 3\text{--}5 \text{ mol mol}^{-1}$ of Rubisco active sites, required for A_{max} in unstressed plants, would decrease A (ref. 11) and may increase concentrations of inhibitors that decrease catalytic activity³⁰. We show that at ~ 2 , 1.5 and $0.5 \text{ mol RuBP mol}^{-1}$ sites, A decreased by 25, 80 and almost 100%, consistent with earlier analysis of water-stressed¹¹ and phosphate-deficient²² plants.

Inhibition of ATP synthesis and large q_{NP} are consistent with continued O_2 evolution and electron transport²⁸. Our measurements show that the reduction state of the pyridine nucleotides in stressed leaves did not alter significantly between $\psi = -0.5$ and -2 MPa, the ratio $\text{NADH+NADPH/NAD+NADP}$ remaining constant at 0.14 (data not shown). Such conditions are damaging to photosynthetic and cellular metabolism¹⁻⁴, with increased production of active oxygen species which damage lipids, proteins and so on^{2,4,28}. In water-stressed leaves, particularly under high solar radiation, decreased A removes the primary sink for reductant and energy^{2,4,18}, although photorespiration dissipates energy^{2-4,27} (however, this is disputed⁷) until A and any internal recycling of CO_2 ceases. Energy dissipation in the chloroplast, for example, by the violaxanthin cycle, may not be sufficient and damage, such as photoinhibition, may ensue^{2-4,28}. Inhibition of ATP synthesis has been rejected (based on indirect assessment of CF_1 activity) in favour of reduced g_{CO_2} as a cause of decreased A in slowly but severely water-stressed field-grown sunflowers^{6,8,19}. Possibly, in the field, adjustment of the root/shoot ratio and $\text{g}_{\text{H}_2\text{O}}$ may precede or be concomitant with the metabolic effects of stress, thus ameliorating or avoiding its impacts¹. Metabolic effects have been ascribed to damage caused by rapid stressing^{4,6,16}, but our studies were on relatively slowly stressed plants and resulted, after several days of severe stress, in irreversible damage and death. The loss of capacity to synthesize ATP, in conjunction with the reductant status, may explain many aspects of metabolism under stress². We suggest, as an example, that inadequate ATP impairs synthesis of proteins³¹, thus endangering cell

functions². However, it may stimulate accumulation of 'heat-shock' (chaperone) proteins, because phosphorylation represses heat-shock factor, which prevents expression of heat-shock proteins³².

We conclude that water stress inhibits photosynthesis through decreased RuBP supply caused by low ATP content, itself a result of decreased ATP synthase, confirming that ATP synthesis is sensitive to cellular dehydration^{9,10}. Decreased ATP content has important consequences for cellular metabolism and is relevant to unravelling the effects of water stress^{2,13,18}. □

Methods

Growth conditions

Sunflower (*Helianthus annuus* L. cultivar Avanti) plants were grown at 25/18 °C day/night temperature, ~70% relative humidity and 360 µmol CO₂ mol⁻¹, under 400 µmol m⁻² s⁻¹ photosynthetically active radiation (PAR) (day length 14 h) in 5 l of soil-based compost and were given ample nutrition and water. The third pair of leaves, attached to the plant, was used for all measurements for two weeks after full expansion.

Water stress treatments

Control plants were fully watered; plants were stressed by replacing part of the water transpired to allow stress to develop over 12 days. Stresses were achieved at different rates by differential watering to ensure that measurements were made on leaves of similar age¹¹. The decrease in ψ was ~0.3 MPa per day in the most severe stress. We measured leaf water potential (ψ) with a Scholander pressure chamber¹¹ on the leaf remaining after freeze-clamping^{11,22,23}, then leaf discs were removed for measurement of RWC¹¹. The correlation between ψ and RWC was $\psi = 5.05 + 0.053\text{RWC}$ ($r^2 = 0.90$). Preliminary studies showed no significant differences between tissue used for gas exchange and the remaining leaf.

Gas exchange measurements

Carbon dioxide exchange and water loss were measured and used to calculate photosynthetic rate, stomatal conductance and sub-stomatal CO₂ concentration (c_i)^{11,21–23}. All measurements were made in triplicate on 10 cm² of intact leaf under 800 µmol m⁻² s⁻¹ PAR, at 25 °C and 1 kPa vapour pressure, in a multi-chamber gas-exchange system, using intact leaves attached to three different plants^{11,22,23}. Measurements on intact leaves at large c_a were made with a Ciras-1 gas analyser (PPSystems).

Stomatal heterogeneity

After gas exchange measurements, leaves were exposed to ¹⁴CO₂ for 10 s and freeze-clamped, followed by autoradiography^{11,14,26}.

Assimilation of CO₂ under oxygen electrode conditions

We measured ¹⁴C accumulation in organic compounds on discs removed from illuminated control and wilted leaves and exposed for 10 min to ¹⁴CO₂ of known specific radioactivity with CO₂ concentrations of up to 0.1 mol mol⁻¹ (10% by volume) in 5 ml glass vials with rubber seals. Conditions were the same as in the oxygen electrode. Leaves were immediately killed and ¹⁴C measured after combustion of the sample.

Equilibrium CO₂ compensation points

Detached leaves were measured in a sealed glass chamber with internal air circulation coupled with an infrared gas analyser (Mark III, ADC); the system had a CO₂ leakage rate less than 0.5 µmol mol⁻¹ h⁻¹ at 800 µmol m⁻² s⁻¹ PAR at 25 °C.

Oxygen evolution

Leaf discs were measured in an oxygen electrode (Hansatech) at 400 µmol m⁻² s⁻¹ PAR, CO₂ concentrations up to 0.1 mol CO₂ mol⁻¹ (refs 3, 4) and 25 °C.

Fluorescence

Chlorophyll *a* fluorescence was measured with a modulated meter (Hansatech), concomitant with gas exchange, for calculation of quenching parameters and F_v/F_m (refs 3, 4, 7, 22, 28).

Biochemical analyses

Leaf areas in the gas exchange chambers were freeze-clamped at liquid nitrogen temperature, reaching 0 °C within 0.05 s and –20 °C in 0.1 s, immediately after gas exchange measurements^{11,22,23,33}. ATP, RuBP, 3-PGA and pyridine nucleotides were measured on 10 cm² leaf extracted in 5% (v/v) perchloric acid as described^{11,16,18,22,23}. We measured Rubisco content and activity as described^{11,33}. The amounts of ATP synthase α - and β -subunits were measured³³: frozen leaf tissue was homogenized in a 50 mM Bicine buffer, pH 7.6, 20 mM EDTA, 1 mM MgCl₂ and 50 µM mercaptoethanol. Material precipitated by centrifugation at 10,000g contained ATP synthase, which was solubilized in a 250 mM Tris buffer, pH 7.6 containing 250 mM dithiothreitol, 10% (w/v) SDS and 10% (w/v) glycerol at an SDS to protein ratio of 4:1. The denatured ATP synthase was separated by electrophoresis and the separated proteins blotted onto PVDF membrane, followed by blocking. The blot was developed using primary polyclonal antibody to the α - and β -subunits of ATP synthase (supplied by J. C. Gray). ATP synthase prepared from sunflower leaves was used as standard to quantify the protein from the experiments by visualizing the luminescence onto radiographic film and scanning. The mass ratio of total ATP synthase to the two subunits was taken as 1.61 (ref. 33).

Received 13 April; accepted 2 September 1999.

1. Lawlor, D. W. & Upreti, D. C. in *Photosynthesis, Photoreactions to Plant Productivity* (eds Abrol, Y. P., Mohanty, P. & Govindjee) 421–445 (Oxford and IBH, New Delhi, 1991).
2. Lawlor, D. W. in *Environment and Plant Metabolism* (ed. Smirnoff, N.) 129–160 (Bios Scientific, Oxford, 1995).
3. Cornic, G. & Massacci, A. in *Photosynthesis and the Environment* (ed. Baker, N. R.) 347–366 (Kluwer, The Netherlands, 1996).
4. Cornic, G. in *Photoinhibition of Photosynthesis* (eds Baker, N. R. & Bowyer, J. R.) 297–313 (Bios Scientific, Oxford, 1994).
5. Evans, L. T. *Feeding the Ten Billion. Plants and Population Growth* (Cambridge Univ. Press, Cambridge, 1998).
6. Ort, D. R., Oxborough, K. & Wise, R. R. in *Photoinhibition of Photosynthesis* (eds Baker, N. R. & Bowyer, J. R.) 315–329 (Bios Scientific, Oxford, 1994).
7. Brestic, M., Cornic, G., Fryer, M. J. & Baker, N. R. Does photorespiration protect the photosynthetic apparatus in French bean leaves from photoinhibition during drought stress? *Planta* **196**, 450–457 (1995).
8. Quick, W. P. *et al.* The effect of water stress on photosynthetic carbon metabolisms in four species grown under field conditions. *Plant Cell Environ.* **15**, 25–35 (1992).
9. Keck, R. W. & Boyer, J. S. Chloroplast response to low leaf water potentials. III. Differing inhibition of electron transport and photophosphorylation. *Plant Physiol.* **53**, 474–479 (1974).
10. Younis, H. M., Boyer, J. S. & Govindjee Conformation and activity of chloroplast coupling factor exposed to low chemical potential of water in cells. *Biochim. Biophys. Acta* **548**, 328–340 (1979).
11. Gimenez, C., Mitchell, V. J. & Lawlor, D. W. Regulation of photosynthetic rate of two sunflower hybrids under water stress. *Plant Physiol.* **98**, 516–524 (1992).
12. Graan, T. & Boyer, J. S. Very high CO₂ partially restores photosynthesis in sunflower at low water potentials. *Planta* **181**, 378–384 (1990).
13. Lauer, M. J. & Boyer, J. S. Internal CO₂ measured directly in leaves. *Plant Physiol.* **98**, 1310–1316 (1992).
14. Pospíšilová, J. & Šantrůček, J. in *Handbook of Photosynthesis* (ed. Pessarakali, M.) 427–441 (Mercel Dekker, New York, 1997).
15. Gunasekera, D. & Berkowitz, G. A. Use of transgenic plants with ribulose 1,5-bisphosphate carboxylase/oxygenase antisense DNA to evaluate the rate limitation of photosynthesis under water stress. *Plant Physiol.* **103**, 629–635 (1993).
16. Sage, R. F., Sharkey, T. D. & Seemann, J. R. The *in vivo* response of the ribulose-1,5-bisphosphate carboxylase activation state and the pool sizes of photosynthetic metabolites to elevated CO₂ in *Phaseolus vulgaris* L. *Planta* **174**, 407–416 (1988).
17. Price, G. D. *et al.* Chloroplast cytochrome *b₆/f* and ATP synthase complexes in tobacco: transformation with antisense RNA against nuclear-encoded transcripts for the Rieske FeS and ATP δ polypeptides. *Aust. J. Plant Physiol.* **22**, 285–297 (1995).
18. Sharkey, T. D. & Badger, M. R. Effects of water stress on photosynthetic electron transport, photophosphorylation and metabolite levels of *Xanthium strumarium* cells. *Planta* **156**, 199–206 (1982).
19. Ortiz-Lopez, A., Ort, D. R. & Boyer, J. S. Photophosphorylation in attached leaves of *Helianthus annuus* at low water potentials. *Plant Physiol.* **96**, 1018–1025 (1991).
20. Haraux, F. & de Kouchkovsky, Y. Energy coupling and ATP synthase. *Photosynthesis Res.* **57**, 231–251 (1998).
21. Boyer, J. S., Chin Wong, S. & Farquhar, G. D. CO₂ and water vapor exchange across leaf cuticle (epidermis) at various water potentials. *Plant Physiol.* **114**, 185–191 (1997).
22. Jacob, J. & Lawlor, D. W. Dependence of photosynthesis of sunflower and maize leaves on phosphate supply, ribulose-1,5-bisphosphate carboxylase/oxygenase activity, and ribulose-1,5-bisphosphate pool size. *Plant Physiol.* **98**, 801–807 (1992).
23. Paul, M. J. *et al.* Reduction in phosphoribulokinase activity by antisense RNA in transgenic tobacco: effect on CO₂ assimilation and growth in low irradiance. *Plant J.* **7**, 535–542 (1995).
24. Laik, A. Calculation of leaf photosynthetic parameters considering the statistical distribution of stomatal apertures. *J. Exp. Bot.* **34**, 1627–1635 (1983).
25. Terashima, I., Wong, S.-C., Osmond, C. B. & Farquhar, G. D. Characterisation of non-uniform photosynthesis induced by abscisic acid in leaves having different mesophyll anatomies. *Plant Cell Physiol.* **29**, 385–394 (1988).
26. Robinson, S. P., Grant, W. J. R. & Loveys, B. R. Stomatal limitations of photosynthesis in abscisic acid-treated and water-stressed leaves measured at elevated CO₂. *Aust. J. Plant Physiol.* **15**, 495–503 (1988).
27. Kozaki, A. & Takeba, G. Photorespiration protects C3 plants from photooxidation. *Nature* **384**, 557–560 (1996).
28. Krause, G. H. & Weis, E. Chlorophyll fluorescence and photosynthesis: The basics. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **42**, 313–349 (1991).
29. Sharp, R. E. & Boyer, J. S. Photosynthesis at low water potentials in sunflower: lack of photoinhibitory effects. *Plant Physiol.* **82**, 90–95 (1986).
30. Edmonson, D. L., Kane, H. J. & Andrews, T. J. Substrate isomerization inhibits ribulosebisphosphate carboxylase-oxygenase during catalysis. *FEBS Lett.* **260**, 62–66 (1990).
31. Riccardi, E., Gazeau, P., de Vienne, D. & Zivy, M. Protein changes in response to progressive water deficit in maize. *Plant Physiol.* **117**, 1253–1263 (1998).
32. Schöffl, F., Prändl, R. & Reindl, A. Regulation of the heat-shock response. *Plant Physiol.* **117**, 1135–1141 (1998).
33. Theobald, J. C., Mitchell, R. A. C., Parry, M. A. J. & Lawlor, D. W. Estimating the excess investment in ribulose-1,5-bisphosphate carboxylase/oxygenase in leaves of spring wheat grown under elevated CO₂. *Plant Physiol.* **118**, 945–955 (1998).

Acknowledgements

We acknowledge help in measurement of metabolites and Rubisco from A. J. Keys, M. A. Parry and M. J. Paul and of ATP-synthase by J. C. Theobald; discussions with R. A. C. Mitchell; and manuscript preparation by K. Lawlor. J. C. Gray supplied the ATP synthase antibody. Financial support to W.T. was from CONICYT, Venezuela and Rothamsted International. IACR-Rothamsted receives grant aid from the BBSRC, UK.

Correspondence and requests for materials should be addressed to D.W.L. (e-mail: david.lawlor@bbsrc.ac.uk).

Neurotrophin-evoked rapid excitation through TrkB receptors

Karl W. Kafitz^{*,†}, Christine R. Rose^{*,†}, Hans Thoenen[‡] & Arthur Konnerth^{*,†}

^{*} Institut für Physiologie, Technische Universität München, Biedersteinerstr. 29, 80802 München, Germany

[†] I. Physiologisches Institut, Universität des Saarlandes, 66421 Homburg, Germany

[‡] Max-Planck-Institut für Neurobiologie, Am Klopferspitz 18A, 82152 Martinsried, Germany

Neurotrophins are a family of structurally related proteins that regulate the survival, differentiation and maintenance of function of different populations of peripheral and central neurons^{1–3}. They are also essential for modulating activity-dependent neuronal plasticity^{4–7}. Here we show that neurotrophins elicit action potentials in central neurons. Even at low concentrations, brain-derived neurotrophic factor (BDNF) excited neurons in the hippocampus, cortex and cerebellum. We found that BDNF and neurotrophin-4/5 depolarized neurons just as rapidly as the neurotransmitter glutamate, even at a more than thousand-fold lower concentration. Neurotrophin-3 produced much smaller responses, and nerve growth factor was ineffective. The neurotrophin-induced depolarization resulted from the activation of a sodium ion conductance which was reversibly blocked by K-252a, a protein kinase blocker which prefers tyrosine kinase Trk receptors⁸. Our results demonstrate a very rapid excitatory action of neurotrophins, placing them among the most potent endogenous neuro-excitants in the mammalian central nervous system described so far.

Figure 1a shows an experiment in which BDNF and glutamate

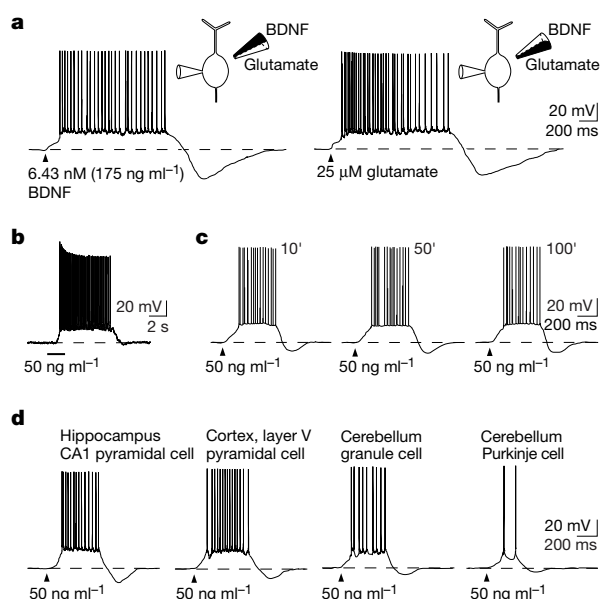


Figure 1 Excitatory action of BDNF in central neurons. **a**, Action potential firing produced by BDNF and glutamate application in a hippocampal CA1 pyramidal neuron. Substances were focally applied to the soma by a double-barrelled stimulation pipette (see Methods). **b**, Train of action potentials produced by bath-applied BDNF. The bar indicates the time of application (2 s). Similar responses were obtained in 7 out of 7 neurons. **c**, Responses evoked by focally applied BDNF at 10, 50 and 100 min after the onset of whole-cell recording. Responses in **b** and **c** were obtained in different CA1 pyramidal neurons. **d**, Comparison of BDNF-evoked responses obtained in different types of central neurons as indicated. Membrane potential was -60 mV unless otherwise indicated; arrowheads show the time of the focal application (40-ms pulse duration).

were alternately pressure ejected from a double-barrelled pipette placed near the cell body of a CA1 pyramidal neuron in a hippocampal slice preparation obtained from juvenile rats (postnatal day 11). BDNF (6.43 nM), at a more than thousand-fold lower concentration than glutamate, had a depolarizing action that was similar to that of glutamate (25 μM) ($n = 6$). A brief (40-ms) pulse-like application of BDNF evoked a rapid depolarization resulting in a train of action potentials (Fig. 1a, left). Ejection of glutamate from the second barrel produced a virtually identical train of action potentials (Fig. 1a, right). The latency of the BDNF-evoked depolarization was also similar to that of glutamate (9.2 ± 0.7 ms and 9.0 ± 0.7 ms, respectively; $n = 6$). This latency includes the time needed for diffusion and the build-up of the effective concentration at the receptor site. The prolonged action of BDNF, considerably exceeding the duration of application, may be attributable at least in part to the time needed for clearance of BDNF from the slice by the bath perfusion. An alternative way of evoking depolarization and action potentials was by bath applying BDNF (Fig. 1b); however, extremely fast perfusion rates ($>5-6$ ml min⁻¹) were required, and the cells were frequently lost after 1–2 applications.

With local application, the BDNF-mediated responses were reproducible for more than 100 minutes, without any noticeable run down (Fig. 1c, $n = 14$). This indicates that the depolarizing action of BDNF may be membrane delimited and does not involve soluble transducer molecules, for example second messengers, which were washed out through the dialysing patch pipette. We tested whether the fast depolarizing action of BDNF also occurred in other brain regions. We used cortical and cerebellar slices from rats of the same age group (postnatal days 9–14) and applied a 'standard' BDNF-pulse (50 ng ml⁻¹, 40-ms duration) to visually identified⁹ cortical layer V pyramidal cells, cerebellar granule cells and Purkinje cells. Although BDNF evoked depolarizing responses

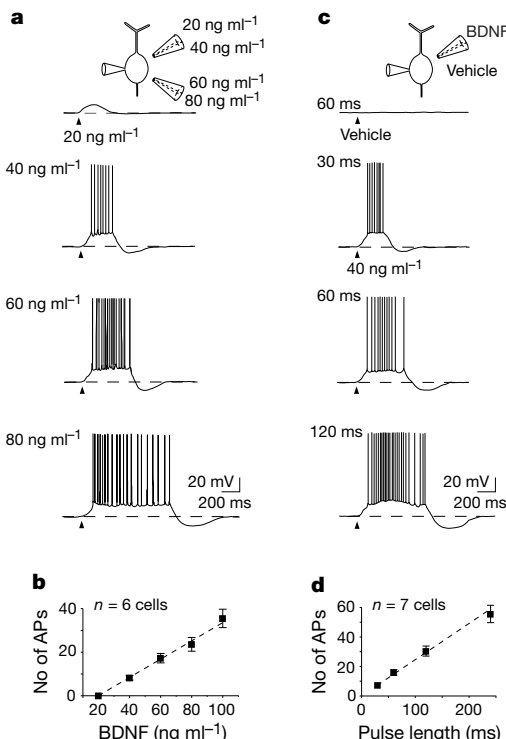


Figure 2 Dose-dependence of BDNF-evoked neuronal firing in hippocampal CA1 pyramidal neurons. **a**, Voltage responses to focal stimulation with increasing concentrations of BDNF (30-ms pulse) applied by a set of two double-barrelled micropipettes. **b**, The mean number of action potentials (APs) $\pm$ s.e.m. versus BDNF concentration. **c**, Responses to focal stimulation with vehicle alone (0.1% BSA in PBS) or to BDNF at increasing pulse durations. **d**, The mean number of APs $\pm$ s.e.m. plotted against application duration of BDNF.

and/or action potential firing in every tested cell of each of these regions ($n = 47$), there were nevertheless cell-type-specific differences in the magnitude of the response (Fig. 1d). In hippocampal ($n = 9$) and cortical pyramidal cells ($n = 9$), similar BDNF-evoked responses were registered (15.7 ± 1.4 and 15.0 ± 1.8 action potentials, respectively). The action of BDNF was weaker in the cerebellum. In cerebellar granule cells, our standard stimulation pulse evoked 8.5 ± 1.6 action potentials ($n = 14$ cells), whereas Purkinje cells responded with just 1.7 ± 0.3 action potentials ($n = 15$).

To establish whether the rapid depolarizing action of BDNF showed a dose–response relationship, we used hippocampal CA1 pyramidal cells and either applied BDNF at different concentrations or changed the pulse duration while keeping the concentration constant. For a pulse of 30-ms duration, the lowest concentration of BDNF used in this series (20 ng ml^{-1}) produced a subthreshold depolarization (Fig. 2a). At increasingly higher concentrations, BDNF elicited trains of action potentials of increasing duration. The number of action potentials was linearly related to the concentrations used (Fig. 2b). When the duration of the application pulse was varied at a constant BDNF concentration (40 ng ml^{-1}), the number of evoked action potentials also increased linearly with the pulse duration (Fig. 2c, d). Furthermore, a dose-dependence of the BDNF action was observed in voltage-clamp experiments (Fig. 3e). Application of vehicle alone was ineffective (Fig. 2c, top). In additional control experiments, we observed that applications of heat-inactivated (3 out of 3) or protease-inactivated (3 out of 3) BDNF were entirely ineffective, confirming the specificity of the BDNF action.

The lowest concentrations of BDNF that caused depolarization were $0.5\text{--}2 \text{ nM}$ ($\approx 13\text{--}50 \text{ ng ml}^{-1}$), and thus are well within the

range found for other BDNF-induced effects requiring activation of TrkB receptors^{10–17}. However, our concentrations refer to those in the application pipettes, and the concentrations at the site of action—namely at the TrkB receptors themselves—are probably lower by a factor of two to four because of local dilution. To analyse whether the depolarizing action of BDNF resulted from an activation of the tyrosine kinase TrkB receptors, we used the protein kinase blocker K-252a which, at the concentration applied (150 nM), shows a relatively high preference for the blockade of the transmembranous tyrosine kinase receptors of the trk gene family⁸. Figure 3a shows a voltage-clamp experiment performed in a hippocampal CA1 pyramidal cell in which three subsequent appli-

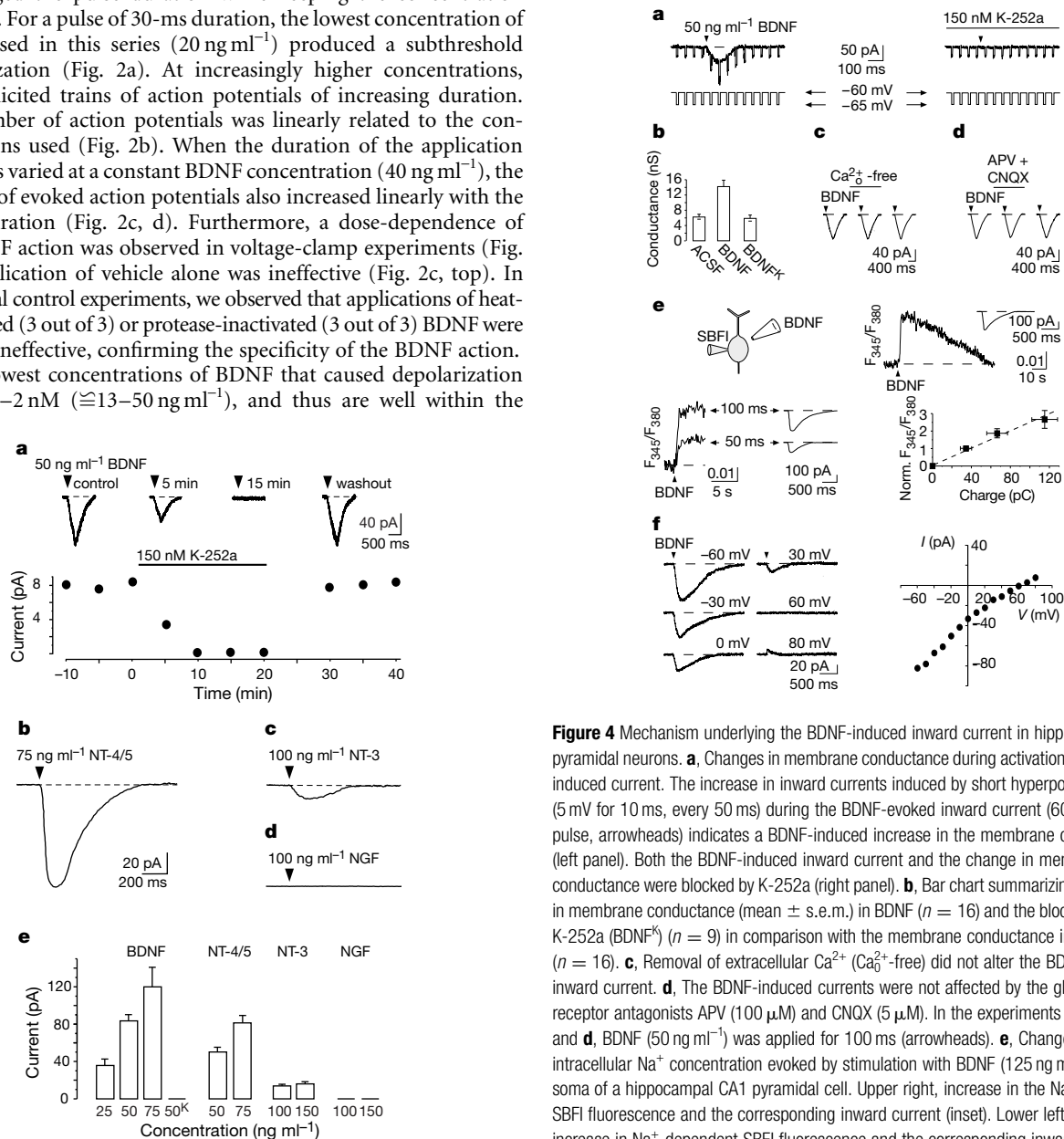


Figure 3 TrkB requirement for the NT-mediated excitation of hippocampal CA1 pyramidal neurons. **a**, Inward currents evoked by BDNF (50 ng ml^{-1} , 100-ms pulse). Top, the BDNF-induced currents are reversibly blocked by the tyrosine kinase blocker K-252a. Bottom, 10 consecutive BDNF-evoked currents before, during and after application of K-252a. Inward currents evoked by focal application of NT-4/5 (100-ms pulse) (**b**) and NT-3 (100-ms pulse) (**c**). NGF (100-ms pulse) did not elicit an inward current (**d**). **e**, Bar graph summarizing the current amplitudes (mean $\pm$ s.e.m.) evoked by BDNF alone ($n = 12$ cells), BDNF in the presence of K-252a (50 ng ml^{-1} ; $n = 8$), NT-4/5 ($n = 10$), NT-3 ($n = 13$) and NGF ($n = 12$) at different concentrations as indicated.

Figure 4 Mechanism underlying the BDNF-induced inward current in hippocampal CA1 pyramidal neurons. **a**, Changes in membrane conductance during activation of the BDNF-induced current. The increase in inward currents induced by short hyperpolarizing steps (5 mV for 10 ms, every 50 ms) during the BDNF-evoked inward current (60-ms BDNF pulse, arrowheads) indicates a BDNF-induced increase in the membrane conductance (left panel). Both the BDNF-induced inward current and the change in membrane conductance were blocked by K-252a (right panel). **b**, Bar chart summarizing the change in membrane conductance (mean $\pm$ s.e.m.) in BDNF ($n = 16$) and the blocking effect of K-252a (BDNF^K) ($n = 9$) in comparison with the membrane conductance in ACSF alone ($n = 16$). **c**, Removal of extracellular Ca²⁺ (Ca²⁺-free) did not alter the BDNF-induced inward current. **d**, The BDNF-induced currents were not affected by the glutamate receptor antagonists APV (100 μM) and CNQX (5 μM). In the experiments depicted in **c** and **d**, BDNF (50 ng ml^{-1}) was applied for 100 ms (arrowheads). **e**, Changes in the intracellular Na⁺ concentration evoked by stimulation with BDNF (125 ng ml^{-1}) in the soma of a hippocampal CA1 pyramidal cell. Upper right, increase in the Na⁺-dependent SBFI fluorescence and the corresponding inward current (inset). Lower left, proportional increase in Na⁺-dependent SBFI fluorescence and the corresponding inward currents in response to the BDNF pulses, as indicated. Lower right, correlation between the BDNF-induced change in relative SBFI fluorescence (Norm. F_{345}/F_{380}) and the evoked change in total charge. Data are expressed as means $\pm$ s.e.m. The first data point represents the application of vehicle alone ($n = 5$). The others were obtained by applying BDNF for 50, 100 and 200 ms respectively. **f**, Current–voltage relationship of the BDNF (50 ng ml^{-1})-evoked currents. Currents were measured at 6 different holding potentials (left panel) and the I–V graph obtained from the same experiment is shown (right panel). Note that the reversal potential of the BDNF-evoked currents was about 60 mV. Similar results were obtained in 4 out of 4 experiments.

cations of 50 ng ml⁻¹ BDNF evoked inward currents. The prolonged application of K-252a for 10–11 min before applying the first BDNF test pulse completely blocked BDNF-elicited inward currents. The block developed gradually, and incubation in K-252a for less than 10 min resulted in an incomplete block. The BDNF-elicited current response returned to control values 10 min after stopping the K-252a administration (Fig. 3a). Similar results were obtained in 8 out of 8 cells. Another tyrosine kinase inhibitor, K-252b (200 nM), which is less effective than K-252a⁸, blocked only 20% of the BDNF-evoked current (BDNF control response: 85.6 ± 9.1 pA; BDNF + K-252b: 69.4 ± 6.5 pA; *n* = 4 cells).

We then considered whether other neurotrophins (NTs) exhibit the same excitatory properties as BDNF. We compared the effect of NT-4/5, NT-3 and nerve growth factor (NGF) on hippocampal CA1 pyramidal cells with that of BDNF. Notably, the most complete information on NT and Trk receptor expression and function is available for the hippocampus^{11,14,15,18–20}. To allow a direct comparison of all four NTs, we used a combination of two double-barrelled application pipettes for each cell (*n* = 6). NT-4/5, which, like BDNF, mediates its biological actions through TrkB receptors^{1,2,21}, elicited inward currents, the amplitudes of which were indistinguishable from those evoked by comparable concentrations of BDNF (Fig. 3b, e). As with BDNF, K-252a completely blocked the NT-4/5-evoked inward currents (*n* = 6; data not shown). In contrast, NT-3, which preferentially binds to TrkC receptors and only cross-reacts weakly with TrkB receptors^{1,2,21}, elicited a much smaller inward current than BDNF or NT-4/5. Even at concentrations of 100–150 ng ml⁻¹, NT-3 produced barely detectable responses (Fig. 3c, e). Lastly, we examined NGF, which virtually exclusively binds to the TrkA receptor^{1,2,21}. In agreement with the absence of TrkA receptors in hippocampal neurons^{10,11,17,18,22,23}, NGF (100–150 ng ml⁻¹) produced no inward current in any of the CA1 pyramidal cells analysed (Fig. 3d, e). Together, our results indicate that the depolarizing actions of NTs in the hippocampus may be mediated by TrkB receptors. The weak excitatory action of NT-3 may reflect a weak cross-activation of the TrkB receptor.

The rapid action of NTs in the millisecond range (Fig. 1a) most likely resulted from the activation of ion channels by TrkB receptors. To test this possibility we applied brief hyperpolarizing pulse-like voltage steps while monitoring BDNF-induced currents. Figure 4a and b shows that the BDNF-induced current was associated with an increase in the amplitude of the brief test currents, demonstrating an increase in membrane conductance. K-252a abolished both the BDNF-induced current and the conductance change. These results clearly indicate that the BDNF-evoked depolarization occurs through activation of ion channels.

In view of the reported BDNF-mediated increase in intracellular Ca²⁺ concentration^{17,20,24}, we investigated whether the inward current was mediated by Ca²⁺ influx. Figure 4c shows that the BDNF-induced current persisted without any detectable change in a Ca²⁺-free perfusion medium, a result that was confirmed in 7 out of 7 cells. We then analysed the role of Na⁺ ions by using ratiometric Na⁺ imaging²⁵, in combination with whole-cell recordings (Fig. 4e). Each BDNF-induced inward current was associated with a transient elevation of the intracellular Na⁺ concentration (*n* = 7). Increasing the inward current by increasing the duration of the BDNF application pulse resulted in a directly proportional increase of both the inflowing charge and the amplitude of the Na⁺-dependent fluorescence transient (Fig. 4e, graph) (*n* = 6). Application of vehicle alone did not elicit any changes in fluorescence nor in charge (*n* = 5). Lastly, we determined the voltage dependence of BDNF-evoked currents (Fig. 4f) and found that the measured reversal potential (59.7 ± 1.87 mV, *n* = 4) is very similar to the calculated equilibrium potential for Na⁺ (58.4 mV). These experiments clearly indicate that the NT-evoked inward current was due to an influx of Na⁺.

Based on evidence indicating that the regulation of glutamate

receptor channels is NT-dependent, these seemed the likely candidates responsible for the depolarizing responses^{14,16}. However, a combination of receptor antagonists of these glutamate-gated ionotropic channels (10 μM 6-cyano-7-nitroquinoxaline-2,3-dione (CNQX) and 100 μM DL-2-amino-5-phosphovaleric acid (APV), *n* = 6 cells) had no effect on the BDNF-induced inward current (Fig. 4d), indicating that glutamate receptor channels did not contribute to the NT-evoked inward current. It also seems unlikely that the NTs acted indirectly, by mediating the release of glutamate (or some other transmitter) from neighbouring cells, because tetrodotoxin (TTX), at a concentration (0.5 μM) at which it completely blocked action potential firing and hence presynaptic transmitter release, had no significant effect on the BDNF-induced inward current (*n* = 5 cells; data not shown). Moreover, this latter experiment also indicates that the depolarizing action of BDNF may be mediated by Na⁺-permeable channels that are not sensitive to TTX.

In conclusion, our results show a new and unexpectedly rapid action of NTs. Although rapid effects of NTs have been reported^{14,16,17,20,24,26}, the time frames in which they came into play were orders of magnitude longer than those reported here. In our experiments, the onset of the BDNF-evoked response could not be distinguished from that of glutamate. The delay of about 9 ms includes the diffusion time and the build-up of the appropriate concentration at the receptor site, as mentioned above. The extremely rapid onset of the response suggests a direct interaction of the transmembranous tyrosine kinase TrkB receptor and an as yet unidentified Na⁺-permeable channel. This is further supported by our results which indicate that this is a membrane-delimited process. The rapidity of the onset of the depolarization in the low millisecond range seems to exclude the possibility that the signal transduction occurs through an activation of the trkB tyrosine kinase activity, as the most rapid biochemically detectable tyrosine phosphorylation of Trk receptors after NT administration is in the range of 30 seconds²⁷. The blocking effect of the tyrosine kinase antagonists K-252a and K-252b does not allow us to determine whether the signal transduction requires a basic level of already phosphorylated TrkB receptors, or whether these compounds block the allosteric activation by binding to the phosphorylation site. This leaves open the issue of whether there is a direct interaction between the TrkB receptor and the Na⁺ channel, or whether interposed adapter molecules are involved. In spite of the lack of details of the underlying mechanisms, the effects of BDNF and NT-4/5 on different populations of central neurons in slice preparations represent an unprecedented rapid action of NTs comparable to that of glutamate. The rapidity of their action, combined with being transported to axon terminals^{28,29}, activating postsynaptic dendrites¹⁷ and being secreted in an activity-dependent manner^{28,30}, assigns to NTs in the low nanomolar range functional properties similar to those of conventional excitatory neurotransmitters. □

Methods

Tissue preparation

Parasagittal hippocampal, cortical and cerebellar slices (250 μm) were prepared from 9–14-day-old Wistar rats using standard procedures³. Following anaesthesia, the animals were decapitated, and the tissue was rapidly removed and placed in ice-cold artificial cerebrospinal fluid (ACSF) containing (in mM): 125 NaCl, 2.5 KCl, 2 CaCl₂, 1 MgCl₂, 26 NaHCO₃, 1.25 NaH₂PO₄, 20 glucose, bubbled with 95% O₂/5% CO₂. After sectioning, slices were kept in ACSF for 30 min at 37 °C and then at 25 °C for up to 7 h. All chemicals were purchased from Sigma unless otherwise stated.

Patch clamp recordings

Somatic whole-cell recordings were performed from cell bodies located near the top surface of slice preparations with an EPC9 amplifier (HEKA) using 'Pulse' software (HEKA) for data acquisition. Pipettes were pulled from borosilicate glass (Hilgenberg) and coated with silicone (RTV 615, GE Silicons). The intracellular solution contained (in mM): 120 K-glucuronate, 10 HEPES, 32 KCl, 4 NaCl, 0.16 EGTA, 4 Mg-ATP, 0.4 GTP, pH 7.3 (current-clamp recordings); or 135 CsCl, 15 NaCl, 0.6 EGTA, 5 TEA, 2 MgCl₂, 10 HEPES, 0.4 GTP, 4 Mg-ATP, pH 7.3 (voltage-clamp recordings). During the experiments, the slices were continuously perfused with ACSF (20–22 °C). Membrane or holding potentials were generally set to –60 mV. All NTs, namely BDNF, NT-4/5, NT-3, and NGF (Gibco), were

diluted in phosphate-buffered saline (PBS) containing 0.1% bovine serum albumin (BSA) and usually applied by a picospritzer (Parker) coupled to standard micropipettes or to double-barrelled theta glass capillaries (Clark). The application pipettes were placed at a distance of roughly 15 μm above the soma. When NTs were applied at intervals of 5 min or longer, the NT-elicited responses remained stable for up to 2 h. At shorter application intervals, probably because of desensitization, the responses were smaller or absent. In some experiments, BDNF was bath-applied. For this purpose it was directly dissolved in ACSF containing (in μM): 10 bicuculline, 5 CNQX, 100 APV. The protein kinase antagonists K-252a or K-252b (both diluted in ACSF; Calbiochem) were either applied to the somata by micropipettes or by a gravitation-based microperfusion system. All plastic and glassware were blocked twice with PBS containing 0.1% BSA before exposing the material to NTs to prevent binding to the storage and application ware. In some control experiments, BDNF was heat inactivated by incubating it for 30 min at 50 °C or protease inactivated by a 15-min treatment with papain (200 U, 0.5 mg ml⁻¹, Boehringer Mannheim).

Sodium imaging

Cells were loaded with the fluorescent Na⁺ indicator dye sodium-binding benzofuran-isophthalate (SBFI, 1 mM, Molecular Probes) through the patch pipette. Fluorescence images were acquired in parallel to the whole-cell recordings by a variable scan digital imaging system (TILL Photonics) attached to an upright microscope (Zeiss Axioskop, 40 $\times$ water immersion objective, NA 0.75). The fluorescence signals from the somata were obtained at excitation wavelengths of 345 nm (isosbestic point) and of 380 nm (Na⁺-sensitive wavelength) and were background corrected. Data were expressed as changes in fluorescence ratio (345/380 nm) using Igor Pro software (Wavemetrics) for analyses.

Received 24 May; accepted 18 August 1999.

- Lewin, G. R. & Barde, Y. A. Physiology of the neurotrophins. *Annu. Rev. Neurosci.* **19**, 289–317 (1996).
- Bothwell, M. Functional interactions of neurotrophins and neurotrophin receptors. *Annu. Rev. Neurosci.* **18**, 223–253 (1995).
- Ibanez, C. F. Emerging themes in structural biology of neurotrophic factors. *Trends Neurosci.* **21**, 438–444 (1998).
- Schuman, E. M. Neurotrophin regulation of synaptic transmission. *Curr. Opin. Neurobiol.* **9**, 105–109 (1999).
- Bonhoeffer, T. Neurotrophins and activity-dependent development of the neocortex. *Curr. Opin. Neurobiol.* **6**, 119–126 (1996).
- Cellerino, A. & Maffei, L. The action of neurotrophins in the development and plasticity of the visual cortex [published erratum in *Prog. Neurobiol.* **50**, 333 (1996)]. *Prog. Neurobiol.* **49**, 53–71 (1996).
- Thoenen, H. Neurotrophins and neuronal plasticity. *Science* **270**, 593–598 (1995).
- Knüsél, B. & Hefti, F. K-252 compounds: modulators of neurotrophin signal transduction. *J. Neurochem.* **59**, 1987–1996 (1992).
- Edwards, F. A., Konnerth, A., Sakmann, B. & Takahashi, T. A thin slice preparation for patch clamp recordings from neurones of the mammalian central nervous system. *Pflügers Arch.* **414**, 600–612 (1989).
- Ip, N. Y., Li, Y., Yancopoulos, G. D. & Lindsay, R. M. Cultured hippocampal neurons show responses to BDNF, NT-3, and NT-4, but not NGF. *J. Neurosci.* **13**, 3394–3405 (1993).
- Kang, H. & Schuman, E. M. Long-lasting neurotrophin-induced enhancement of synaptic transmission in the adult hippocampus. *Science* **267**, 1658–1662 (1995).
- Figurov, A., Pozzo-Miller, L. D., Olafsson, P., Wang, T. & Lu, B. Regulation of synaptic responses to high-frequency stimulation and LTP by neurotrophins in the hippocampus. *Nature* **381**, 706–709 (1996).
- Lessmann, V., Gottmann, K. & Heumann, R. BDNF and NT-4/5 enhance glutamatergic synaptic transmission in cultured hippocampal neurones. *Neuroreport* **6**, 21–25 (1994).
- Levine, E. S., Crozier, R. A., Black, I. B. & Plummer, M. R. Brain-derived neurotrophic factor modulates hippocampal synaptic transmission by increasing N-methyl-D-aspartate receptor activity. *Proc. Natl Acad. Sci. USA* **95**, 10235–10239 (1998).
- Patterson, S. L. et al. Recombinant BDNF rescues deficits in basal synaptic transmission and hippocampal LTP in BDNF knockout mice. *Neuron* **16**, 1137–1145 (1996).
- Suen, P. C. et al. Brain-derived neurotrophic factor rapidly enhances phosphorylation of the postsynaptic N-methyl-D-aspartate receptor subunit 1. *Proc. Natl Acad. Sci. USA* **94**, 8191–8195 (1997).
- Canossa, M. et al. Neurotrophin release by neurotrophins: implications for activity-dependent neuronal plasticity. *Proc. Natl Acad. Sci. USA* **94**, 13279–13286 (1997).
- Ringstedt, T., Lagercrantz, H. & Persson, H. Expression of members of the trk family in the developing postnatal rat brain. *Brain Res. Dev.* **72**, 119–131 (1993).
- Yan, Q. et al. Immunocytochemical localization of TrkB in the central nervous system of the adult rat [published erratum in *J. Comp. Neurol.* **382**, 546–547 (1997)]. *J. Comp. Neurol.* **378**, 135–157 (1997).
- Berninger, B., Garcia, D. E., Inagaki, N., Hahnel, C. & Lindholm, D. BDNF and NT-3 induce intracellular Ca²⁺ elevation in hippocampal neurones. *Neuroreport* **4**, 1303–1306 (1993).
- Barbacid, M. Neurotrophic factors and their receptors. *Curr. Opin. Cell Biol.* **7**, 148–155 (1995).
- Merlio, J. P., Ernfrors, P., Jaber, M. & Persson, H. Molecular cloning of rat trkC and distribution of cells expressing messenger RNAs for members of the trk family in the rat central nervous system. *Neuroscience* **51**, 513–532 (1992).
- Korte, M. et al. Virus-mediated gene transfer into hippocampal CA1 region restores long-term potentiation in brain-derived neurotrophic factor mutant mice. *Proc. Natl Acad. Sci. USA* **93**, 12547–12552 (1996).
- Berninger, B. & Poo, M. Fast actions of neurotrophic factors. *Curr. Opin. Neurobiol.* **6**, 324–330 (1996).
- Rose, C. R. & Ransom, B. R. Regulation of intracellular sodium in cultured rat hippocampal neurones. *J. Physiol. (Lond.)* **499**, 573–587 (1997).
- Lohof, A. M., Ip, N. Y. & Poo, M. M. Potentiation of developing neuromuscular synapses by the neurotrophins NT-3 and BDNF. *Nature* **363**, 350–353 (1993).

- Kahle, P., Barker, P. A., Shooter, E. M. & Hertel, C. p75 nerve growth factor receptor modulates p140trkA kinase activity, but not ligand internalization, in PC12 cells. *J. Neurosci. Res.* **38**, 599–606 (1994).
- Blöchl, A. & Thoenen, H. Localization of cellular storage compartments and sites of constitutive and activity-dependent release of nerve growth factor (NGF) in primary cultures of hippocampal neurons. *Mol. Cell Neurosci.* **7**, 173–190 (1996).
- Zhou, X. F. & Rush, R. A. Endogenous brain-derived neurotrophic factor is anterogradely transported in primary sensory neurons. *Neuroscience* **74**, 945–953 (1996).
- Goodman, L. J. et al. Regulated release and polarized localization of brain-derived neurotrophic factor in hippocampal neurons. *Mol. Cell Neurosci.* **7**, 222–238 (1996).

Acknowledgements

We thank R. Trautmann and E. Eilers for expert technical assistance and M. Ashdown for editorial assistance. This study was supported by a fellowship from the DFG to K.W.K. and by grants from the Deutsche Forschungsgemeinschaft and the Human Frontier Science Program to A.K.

Correspondence and requests for materials should be addressed to A.K. (akonnerth@med-rz.uni-sb.de).

Complete sequence and gene map of a human major histocompatibility complex

The MHC sequencing consortium*

Here we report the first complete sequence and gene map of a human major histocompatibility complex (MHC), a region on chromosome 6 which is essential to the immune system (reviewed in ref. 1). When it was discovered over 50 years ago the region was thought to specify histocompatibility genes, but their nature has been resolved only in the last two decades. Although many of the 224 identified gene loci (128 predicted to be expressed) are still of unknown function, we estimate that about 40% of the expressed genes have immune system function. Over 50% of the MHC has been sequenced twice, in different haplotypes, giving insight into the extraordinary polymorphism and evolution of this region. Several genes, particularly of the MHC class II and III regions, can be traced by sequence similarity and synteny to over 700 million years ago, clearly predating the emergence of the adaptive immune system some 400 million years ago. The sequence is expected to be invaluable for the identification of many common disease loci. In the past, the search for these loci has been hampered by the complexity of high gene density and linkage disequilibrium.

The impetus to obtain the complete sequence of the MHC was provided by the complex biology and genetics of the histocompatibility regions H-2 and HLA^A. It is therefore no surprise that the MHC at 6p21.31 is among the first multi-megabase regions of the human genome to be completely sequenced. With over 200 identified loci, the MHC is the most gene-dense region of the human genome sequenced so far. It also encodes the most polymorphic human proteins, the class I and class II molecules, some of which have over 200 allelic variants. This extreme polymorphism is thought to be driven and maintained by the long-standing battle for supremacy between our immune system and infectious pathogens. Underlining its biomedical importance, the MHC is associated with more diseases than any other region of the human

* Corresponding authors representing the four sequencing centres: S. Beck (The Sanger Centre, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK); D. Geraghty (Fred Hutchinson Cancer Research Center, Seattle, Washington 98104, USA); H. Inoko (Tokai University School of Medicine, Department of Molecular Life Science, Bohseidai, Isehara, Kanagawa 259-11, Japan); and L. Rowen (University of Washington, Seattle, Washington 98195, USA). A full list of contributors appears at the end of the paper.

genome, including most, if not all, autoimmune conditions (for example, rheumatoid arthritis and diabetes²). Phenotypes with different aetiologies have also been linked to the region, ranging from cancer to sleeping and reading disorders.

The 3.6-megabase (Mb) virtual MHC sequence reported here is derived from a patchwork of different haplotypes and has been assembled from the work of four main groups³ (see Acknowledgements). In addition to the gene map described below, further interesting attributes have started to emerge from preliminary analyses. Long before the term 'single nucleotide polymorphism' (SNP) was coined, variations in the coding and noncoding sequences were observed in the MHC of different individuals and were used to delineate ancestral haplotypes⁴. The available data already indicate that the extreme polymorphism characterizing the MHC is not homogeneous throughout the entire region. The variation in the noncoding sequences appears to peak around the most polymorphic gene loci and 'hitch-hiking', as the result of overdominant allele selection (heterozygote advantage), has been suggested to explain this phenomenon. Variation levels of 5–17% have been reported at some of these loci (HLA-DP, DQ, B and C), which are by far the highest levels found in the human genome so far^{5–8}. A systematic analysis and verification of SNPs across the entire MHC is still in progress. Sequence comparisons revealed several MHC genes (TUBB, TNXB, PBX2, NOTCH4, RXRB and RPS18) to be syntenic in invertebrate genomes such as *Drosophila melanogaster* and *Caenorhabditis elegans*, indicating that the origin of the locus now known as MHC predates the emergence of the adaptive immune system⁹.

Currently, the MHC is the second longest contiguous sequence in the human genome. The determination of very long sequences will allow the exploration of global chromosomal features such as isochores (long-range regions of homogenous G + C content), replicons and repeat dynamics in much greater detail than before. The low G + C isochore covering the classical class II region (see poster or Supplementary Information) is one of the best studied isochores in the human genome^{10,11}. Its predicted boundaries correlate precisely with switching of replication timing from 'later' replication in the classical class II region to 'earlier' replication at the centromeric boundary (P. Jonhonnott and D. Sheer, personal communication) and at the telomeric boundary¹². This may represent a link between isochores and the elusive replicon structure of the human genome^{13,14}.

Figure 1 and the poster accompanying this issue of *Nature* (also available as Supplementary Information) show the complete gene map of all 224 gene loci identified in this particular composite of MHC haplotypes. A list of the genes and their alternative designations is provided as Supplementary Information. Ninety-three of the 224 loci (41.5%) were discovered or located to the MHC as a direct result of the genomic sequence. Historically, the MHC has been divided into three regions: class II (centromeric), class III and class I (telomeric)¹⁵. Analyses of the immediate flanking regions reveal that the 'classical' class I and class II regions extend much further than previously thought^{11,16}. These regions are here referred to as 'extended' class I and class II regions. A set of more than seven genes involved in inflammation, including three members of the tumour necrosis factor (TNF) superfamily, within the class III region is sometimes specified as the class IV region¹⁷. Various other genes associated with the immune system are distributed throughout the MHC. The total immune system constituent of the MHC is 39.8% of the expressed loci (see poster or Supplementary Information) including at least 10 novel genes identified from the genomic sequence (Fig. 1). The clustering of immune-related genes in the MHC region is so striking that it seems unlikely to be coincidental¹⁸. The classical class II region is particularly notable because almost all of the genes are of immune function, namely class II A and B genes, LMPs, TAPs and TAPBP in the extended class II region. This clustering of immunity genes in the MHC may reflect

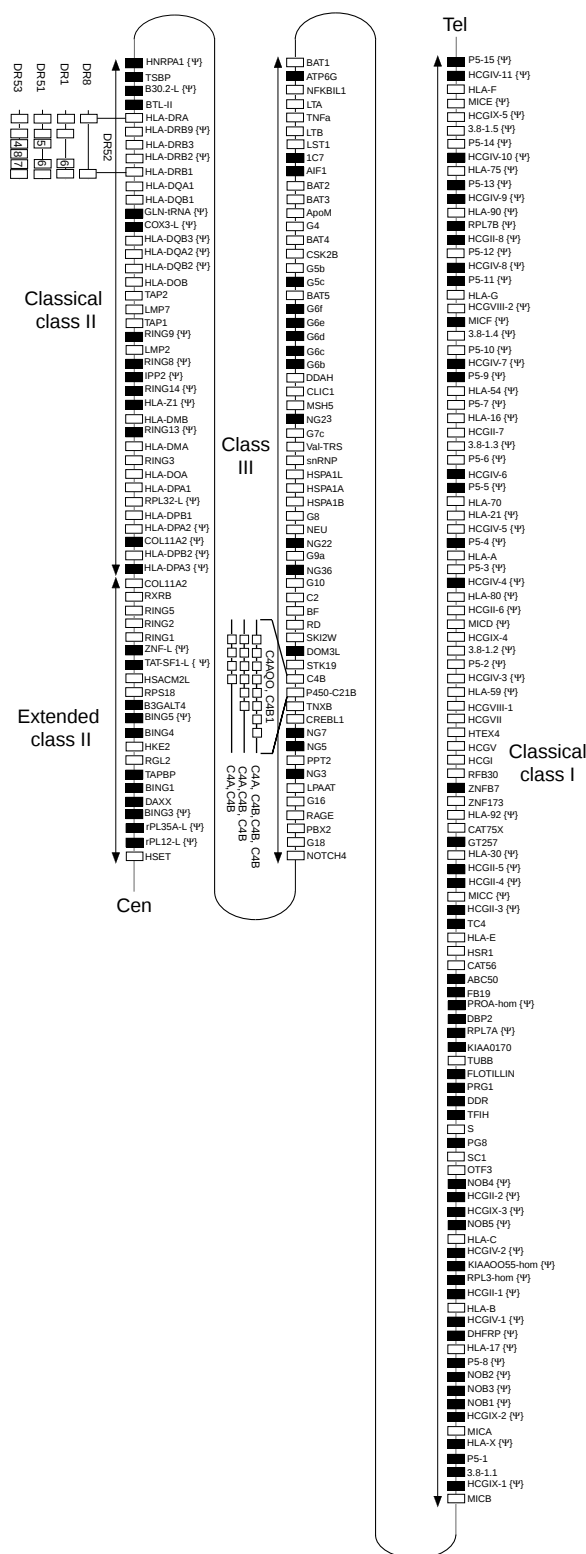


Figure 1 Complete gene map of the MHC reference sequence reported here. Genes are shown in order from telomere to centromere but not to scale. Gene loci that were discovered or located to the MHC as a direct result of the genomic sequence are indicated by filled boxes. As will be the case for the rest of the human genome, the MHC reference sequence is a composite of different haplotypes. However, in regions with known differences in gene content (C4 region in class III and DR region in the classical class II region) only single haplotypes were sequenced (C4A0Q, C4B1 in the class III region and DR52 in the class II region).

co-evolution of functions or co-expression of related transcripts. However, the proportion of immune system genes in the genome in general is not known at present and it may be equally high. The average gene density (including pseudogenes) over the entire 3.6 Mb of the MHC is 1 gene per 16 kilobases (kb), with distinct regional variations. Particularly interesting is the proportion of expressed genes in relation to pseudogenes. Except in certain haplotypes, where the C4 regions have duplicated (Fig. 1), there appear to be no or very few pseudogenes in the class III region. In contrast, the class I and class II regions are full of pseudogenes. Both class I and class II regions appear to have duplicated many times, generating novel gene family members which have then diverged into new functions. A possible explanation for maintaining such high levels of pseudogenes could be that they are involved in generating new alleles by gene conversion, a phenomenon that has been observed at other human immune loci¹⁹.

The first complete MHC sequence provides an important new tool for studying the genetics, biology and evolution of human multigene families, populations and disease. In the long term, the biological importance of the MHC is likely to justify the re-sequencing and epigenetic analysis of several common haplotypes, which differ markedly in sequence and gene content. Efforts towards this goal are already in progress, as they will facilitate the precise identification of many disease loci. Typing of new micro-satellites derived from the genomic sequence has already allowed us to narrow down the candidate region for psoriasis vulgaris (an inflammatory skin disorder) to a critical segment including four genes—S, PG8, SC1 and OTF3 (H. Inoko, unpublished data). The discovery of regions paralogous to the MHC on chromosomes 1, 9 and 19 is indicative of ancient duplications and possibly the remnant of tetraploidization in the early vertebrate genome^{20,21}. The sequences of these related regions and of other vertebrate MHCs are eagerly awaited²². □

Received 20 July; accepted 26 August 1999.

- Klein, J. *Natural History of the Major Histocompatibility Complex* (Wiley, New York, 1986).
- Tiwari, J. L. & Terasaki, P. I. *HLA and Disease Associations* (Springer, Berlin, 1985).
- <http://www.sanger.ac.uk/HGP/Chr6/MHC.shtml>
- Dawkins, R. *et al.* Genomics of the major histocompatibility complex: haplotypes, duplication, retroviruses and disease. *Immunol. Rev.* **167**, 275–304 (1999).
- Limm, T. M., Ashdown, M. L., Naughton, M. J., McGinnis, M. D. & Simons, M. D. HLA-DQA1 allele and suballele typing using non-coding sequence polymorphisms. *Hum. Immunol.* **38**, 57–68 (1993).
- Horton, R. *et al.* Large-scale sequence comparisons reveal unusually high levels of variation in the HLA-DQB1 locus in the class II region of the human MHC. *J. Mol. Biol.* **282**, 71–97 (1998).
- Guillaudoux, T., Janer, M., Wong, G. K., Spies, T. & Geraghty, D. E. The complete genomic sequence of 424,015 bp at the centromeric end of the HLA class I region: gene content and polymorphism. *Proc. Natl Acad. Sci. USA* **95**, 9494–9499 (1998).
- Satta, Y., Kupfermann, H., Li, Y.-J. & Takahata, N. Molecular clock and recombination in primate MHC genes. *Immunol. Rev.* **167**, 367–379 (1999).
- Trachtulec, Z. *et al.* Linkage of TATA-binding protein and proteasome subunit C5 genes in mice and humans reveals synteny conserved between mammals and invertebrates. *Genomics* **44**, 1–7 (1997).
- Fukagawa, T. *et al.* A boundary of long-range G+C% mosaic domains in the human MHC locus: pseudoautosomal boundary-like sequence exists near the boundary. *Genomics* **25**, 184–191 (1995).
- Stephens, R. *et al.* Gene organisation, sequence variation and isochore structure at the centromeric boundary of the human MHC. *J. Mol. Biol.* **291**, 789–799 (1999).
- Tenzen, T. *et al.* Precise switching of DNA replication timing in the GC content transition area in the human MHC. *Mol. Cell. Biol.* **17**, 4043–4050 (1997).
- Hamlin, J. L. & Dijkwel, P. A. On the nature of replication origins in higher eukaryotes. *Curr. Opin. Genet. Dev.* **5**, 153–161 (1995).
- Delgado, S., Gomez, M., Bird, A. & Antequera, F. Initiation of DNA replication at CpG islands in mammalian chromosomes. *EMBO J.* **17**, 2426–2435 (1998).
- Trowsdale, J. & Campbell, R. D. Map of the human major histocompatibility complex. *Immunol. Today* **18** (Suppl.) (1997).
- Ruddy, D. A. *et al.* A 1.1-Mb transcript map of the hereditary hemochromatosis locus. *Genome Res.* **7**, 441–456 (1997).
- Gruen, J. R. & Weissman, S. M. Evolving views of the major histocompatibility complex. *Blood* **90**, 4252–4256 (1997).
- Bodmer, W. F. Evolutionary significance of the HLA system. *Nature* **237**, 139–145 (1997).
- Haino, M. *et al.* Comparison and evolution of human immunoglobulin VH segments located in the 3' 0.8-megabase region: Evidence for unidirectional transfer of segmental gene sequences. *J. Biol. Chem.* **269**, 2619–2626 (1994).
- Sugaya, K. *et al.* Three genes in the human MHC class III region near the junction with the class II: Gene for receptor of advanced glycosylation end products, PBX homeobox genes and a Notch homolog, human counterpart of mouse mammary tumor gene int-3. *Genomics* **23**, 408–419 (1994).
- Kasahara, M. The chromosomal duplication model of the major histocompatibility complex. *Immunol. Rev.* **167**, 17–32 (1999).

22. Kaufman, J. *et al.* The chicken B locus is a minimal essential major histocompatibility complex. *Nature* (this issue).

Supplementary Information is available at Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

Sequencing of the class I region in the US was funded by a Department of Energy grant to M. Olson and by US NIH grants to D.G. Sequencing of the class I region in Japan was funded by the Japanese Science and Technology Corporation (JST). Sequencing of the class II region at the Sanger Centre was funded by the Wellcome Trust. Early stages of the class II sequence were funded by the Imperial Cancer Research Fund and the EU BioMed1 programme. Sequencing of the class III region at the University of Washington Multimegabase Sequencing Center was funded by the US NIH and Department of Energy. We thank all members past and present of the University of Washington Genome Center (<http://www.genome.washington.edu/uwgc/>), the Genome Sequencing Group at Tokai University (<http://www.alis.tokyo.jst.go.jp/HGS/top.html>), the Chromosome 6 Project Group at the Sanger Centre (<http://www.sanger.ac.uk/HGP/Chr6>) and the University of Washington Multimegabase Sequencing Center (http://chroma.mbt.washington.edu/msg_www/) for their contributions. R.D.C. was funded by the UK Medical Research Council and J.T. was funded by the Wellcome Trust. The work was divided as follows: The Sanger Centre, Hinxton, UK: HSET to TSBP, 1.2 Mb; Tokai University, Isehara, Japan: HSPA1B to P5-15, 2.2 Mb; Fred Hutchinson Cancer Research Center and the University of Washington Genome Center (UWGC), Seattle, USA: NFKB1 to HLA-F and beyond, 2.2 Mb; University of Washington Multimegabase Sequencing Center (UWMS), Seattle, USA: NOTCH4 to LTA, 0.76 Mb. From these sequences a 3,673,800-base-long consensus was assembled covering P5-15 to HSET (http://www.sanger.ac.uk/HGP/Chr6/MHC_990719.fasta).

Correspondence and requests for materials should be addressed to D. Geraghty (e-mail: geraghty@fhcr.org) or H. Inoko (e-mail: hinoko@is.icc.u-tokai.ac.jp) for the class I region, S. Beck (e-mail: beck@sanger.ac.uk) or J. Trowsdale (e-mail: jt233@mole.bio.cam.ac.uk) for the class II region and D. Campbell (e-mail: rcampbel@hgmp.mrc.ac.uk) or L. Rowen (e-mail: leerowen@u.washington.edu) for the class III region.

Contributors in alphabetical order (numbers in parentheses indicate addresses): B. Aguado (5), S. Bahram (7), S. Beck (1), R. D. Campbell (5), S. A. Forbes (6), D. Geraghty (2), T. Guillaudoux (2), L. Hood (4), R. Horton (1), H. Inoko (3), M. Janer (2), C. Jasoni (2), A. Madan (4), S. Milne (1), M. Neville (5), A. Oka (3), S. Qin (4), G. Ribas-Despuig (5), J. Rogers (1), L. Rowen (4), T. Shiina (3), T. Spies (2), G. Tamiya (3), H. Tashiro (8), J. Trowsdale (6), Q. Vu (2), L. Williams (2), M. Yamazaki (8) (1) The Sanger Centre, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK (2) Fred Hutchinson Cancer Research Center, Seattle, Washington 98104, USA (3) Tokai University School of Medicine, Department of Molecular Life Science, Bohseidai, Isehara, Kanagawa 259-11, Japan (4) University of Washington, Seattle, Washington 98195, USA (5) HGMP Resource Centre, Wellcome Trust Genome Campus, Hinxton CB10 1SB, UK (6) Cambridge University, Department of Pathology, Division of Immunology, Tennis Court Road, Cambridge CB2 1QP, UK (7) Centre de Recherche d'Immunologie et d'Hématologie, 4 Rue Kirschleger, 67085 Strasbourg, France (8) Bioscience Research Laboratory, Fujiya Co., Ltd., 228 Soya, Hadano, Kanagawa 257, Japan

The chicken B locus is a minimal essential major histocompatibility complex

Jim Kaufman*, Sarah Milne†, Thomas W. F. Göbel‡, Brian A. Walker*, Jansen P. Jacob*, Charles Auffray§, Rima Zoorob§ & Stephan Beck†

* Institute for Animal Health, Compton RG20 7NN, UK

† The Sanger Centre, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK

‡ Institute for Animal Physiology, Munich 80539, Germany

§ CNRS, Genetique Moléculaire et Biologie du Développement, UPR420, Villejuif Cedex, France

Here we report the sequence of the region that determines rapid allograft rejection in chickens, the chicken major histocompatibility complex (MHC). This 92-kilobase region of the B locus^{1–4} contains only 19 genes, making the chicken MHC roughly 20-fold smaller than the human MHC⁵. Virtually all the genes have counterparts in the human MHC, defining a minimal essential set of MHC genes conserved over 200 million years of divergence between birds and mammals. They are organized differently, with

the class III region genes located outside the class II and class I region genes. The absence of proteasome genes^{5,6} is unexpected and might explain unusual peptide-binding specificities of chicken class I molecules. The presence of putative natural killer receptor gene(s)^{5,7} is unprecedented and might explain the importance of the B locus in the response to the herpes virus responsible for Marek's disease⁸⁻¹⁰. The small size and simplicity of the chicken MHC allows co-evolution of genes as haplotypes over considerable periods of time, and makes it possible to study the striking MHC-determined pathogen-specific disease resistance⁸⁻¹⁰ at the molecular level.

There are two loci with class I and class II genes in chickens, the B-F/B-L region of the B locus and the Rfp-Y locus, which are genetically unlinked^{3,4,11}. The B-F/B-L region (but not the Rfp-Y locus) has all of the signal attributes of the MHC of well studied mammals: it contains the classical class I and class II genes; and it determines serological alloantigens, rapid allograft rejection, strong mixed lymphocyte reaction and cellular cooperation in the immune response^{1-4,12-14}. The B-F/B-L region also strongly determines resistance and susceptibility to certain infectious pathogens⁸⁻¹⁰ and we sequenced this region to understand these disease associations.

The B-F/B-L region is simple and compact, with 19 genes found in 92 kilobases (kb) of DNA (Fig. 1). The central portion, stretching from the class II β to the class I genes, is especially compact with no repetitive elements or repeats. These eleven genes in roughly 44 kb have average intron sizes of 200 nucleotides and intergenic distances (excluding promoters) as small as 30 nucleotides, resulting in genes that are as little as one-third the size of their mammalian homologues. There is little change in base composition (roughly 60% G + C) over this region, providing no evidence for the striking isochore structure that correlates with the class I, II and III regions

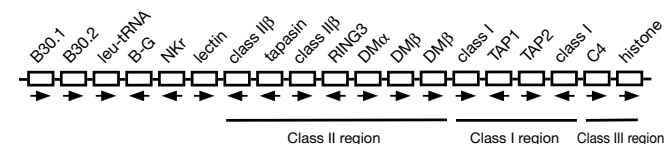


Figure 1 The order of genes in the B-F/B-L region from the B12 haplotype. Arrows indicate transcriptional orientation. The annotated sequence is deposited in the EMBL gene bank, accession number AL023516.

a

```

MD.....EEITYADLRHPTGSLPPAKRQR.VWALPVWDFQK
--NQGVIIYSDLNLPNPKRQQRKPKGNKSSILATEQ----E-NLQKA-QDFQGNKTYHCKDL-SAPE-

ACMLGVLLLLLLLLVTGLSVCVFQKTPPPSAARHCPETNGRNGTELCESTIEYQFCQSTWNSSAAPAA
.LIV-I-GIIC-I-MASVVTI-VIPS-LIQ.....NSSLNTRTQKARH.....

! ! ! ! !
CLLCPQFWRLGLDRCYELSTEGNWTQAKMKCENLQSQLAVLRKKAEDHLQKMAGAEPVWIGLEVSTNQW
-GH--EE-ITYSNS--YIGK-RRT-EESLLA-TSKN-S-LSI.DNE--MKFLSIISPS.---VFRNSSHH

! ! ! ! !
KWVDNSS...YNSTESDNLVSMENRCGTFKNTKVEGDVCSGEHQWVCQKE
P--TMNGLAFKHEIKD-----A-LN-AVLQVNRKLSAQ-GSSIIYH-KHK
  
```

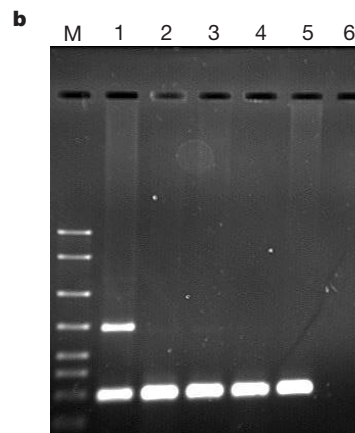


Figure 2 The Nkr (B-NK1) gene encodes an NK receptor. **a**, Predicted amino-acid sequence of cDNA from a chicken natural killer cell line amplified using primers based on the gene B-NK1 (above; deposited in EMBL gene bank, accession number aj245903, compared with mouse NKPR1 (below, accession number p26715). Asterisks, ITIM; overline, transmembrane region; exclamation marks, highly conserved residues in C-type

of the human MHC⁵.

Orthologues of some mammalian MHC genes are found in the B-F/B-L region, but many are absent. There are classical class I genes and TAP genes, but no proteasome (LMP) genes in the vicinity, unlike the mammalian MHC in which LMP genes are intermingled with the TAP genes⁵. Classical class II β genes as well as DM α and DM β genes are present (along with the RING3 gene¹⁵ that encodes a nuclear kinase), but not classical class II α genes or DN/DZ α and DO β genes. In fact, the single nonpolymorphic classical class II α gene is located 5 centimorgans (cM) away from the B locus¹⁶, in contrast to the mammalian MHC in which closely linked α/β gene pairs encode the expressed heterodimeric glycoproteins. There is only one gene of the class III region, the complement component C4. Genes similar to the chicken histone, leu-tRNA, tapasin and B30 genes are found in or just centromeric to the human MHC⁵. Thus, most of the genes present in the B-F/B-L region have homologues (or orthologues) in the mammalian MHC, defining a minimal set of genes conserved over evolutionary time.

The B-F/B-L region is not organized in the same way as the MHC of mammals. The chicken TAP genes are flanked by the class I genes, rather than being located in the class II region⁵. Also, the chicken tapasin gene¹⁷ is located between the two class II β genes, compared with being 100-kb centromeric to the class II β genes in humans and mice⁵. Finally, the remnants of the three regions are in a different order: the chicken class III region (represented by the C4 gene) is located outside the class II region (containing class II β , RING3 and DM genes) and the class I region (containing the class I and the TAP genes). This last difference in organization is an important reason why the chicken MHC is so compact: the highly polymorphic class I and class II genes responsible for graft rejection are not separated by a huge class III region. Indeed, it seems likely that the order of the regions in the chicken MHC is primordial¹⁸, indicating that mammalian MHCs arose by rearrangement.

Some genes in the B-F/B-L region have never been reported to be located in an MHC. The B-G gene is unique to chicken, although it is distantly related to butyrophilin genes found in the mammalian MHC^{5,19,20}. However, most striking is the presence of two genes containing exons encoding C-type animal lectins. Both genes encode type II transmembrane proteins with apparent immunoreceptor tyrosine-based inhibitory motifs (ITIMs) in the amino-terminal cytoplasmic tail. One is closely related to lectin-like natural killer (NK) receptors⁷ and is transcribed in chicken NK cell lines but not in other haemopoietic cell types examined (Fig. 2), identifying it

as the first non-mammalian NK receptor gene reported to our knowledge.

The chicken MHC differs from the MHC of typical mammals in certain functional phenomena, particularly the strong association with resistance and susceptibility to certain infectious pathogens^{8–10}. The fact that the B-F/B-L region is simple and compact compared with that of mammals, with many expected genes absent and a few unexpected genes present, might explain some of these functional associations at a molecular level. For example, LMP genes are not found in the sequenced region (and may be deleted from the chicken genome altogether²¹) which might explain why some chicken classical class I molecules bear peptides with acidic residues at the carboxy terminus⁴. In contrast, mammalian class I molecules only bear peptides that end in hydrophobic or basic amino acids²², and this property has been ascribed to the LMP2 and LMP7 genes, which are reported to shift the proteolytic specificity of the proteasome from acidic residues to hydrophobic and basic residues⁶.

As another example, the presence of NK receptor gene(s) indicates that the chicken MHC may determine responsiveness of certain NK cells, which could be the basis for the unexplained association of the chicken MHC with the response to the herpes virus that causes Marek's disease. The mouse NK receptor locus determines resistance to herpes viruses⁷, and, in fact, the other important (non-MHC) locus that determines resistance to Marek's disease in chickens is syntenic with the mouse NK receptor locus²³. NK cells in mammals detect differences in cell surface expression of class I molecules⁷. We have found that chicken MHC haplotypes determine striking differences in cell surface expression of class I molecules, with the rank order of MHC haplotypes for cell surface class I expression being the same as MHC-determined susceptibility to Marek's disease⁴. We predict, therefore, that chicken class I alleles will be specifically recognized by the NK receptor allele(s) in the B-F/B-L region, although the detailed mechanism is not yet clear.

This last point illustrates perhaps the most important implication of the simple and compact nature of the B-F/B-L region: a low level of recombination, so that genes within the chicken MHC can co-evolve, giving rise to distinct allelic combinations (or haplotypes) that are relatively stable in evolution. In fact, not a single recombinant has been found between the genes determining the serologically detected class I and class II molecules in over 6,000 progeny²⁴. Therefore, the class I, TAP, tapasin and NK receptor genes as well as the class IIβ and DM genes in the chicken MHC can co-evolve as polymorphic genes over a considerable span of time, each combination of alleles manifesting a particular phenotype.

That such co-evolution can have functional consequences is clear from studies with mouse class II genes and with mammalian class I and TAP genes^{25–27}. In these cases, pairs of genes that are relatively closely linked co-evolve so that their products can interact, but genes that are more separated by recombination require one partner to be non-polymorphic, evolving to a 'best average fit' for the alleles of the polymorphic genes which switch frequently by recombination. Chicken TAP and classical class I genes are only tens of nucleotides apart and are virtually inseparable, which should lead to very tight co-evolution. In the extreme case, only one and the same peptide-translocation and peptide-binding specificity would be present in each haplotype²⁷. In fact, we find high polymorphism in chicken TAP genes, as well as strong expression of only one chicken class I gene, whose peptide-binding specificity determines resistance and susceptibility to Rous sarcoma virus⁴ (B.A.W., A. Hoffmann, K. Hala and J.K., unpublished data). A similar picture is emerging for class IIβ genes and antibody responses to killed viral vaccines^{28,29}. Compared with the mammalian MHC, the simple and compact structure of the chicken MHC leads to stronger associations with resistance and susceptibility to infectious pathogens, which suggests that the chicken MHC will be a simpler model with which to understand the molecular basis of such disease susceptibility. □

Methods

The cosmids c4.5, cβ12 and cBF23 (ref. 3) were subjected to shotgun sequencing using fluorescent dye primers on automated sequencers. Gene location and orientation were determined by sequence similarity to mammalian homologues and by comparison with chicken complementary DNA clones (unpublished data).

CD8⁺CD3⁺ cells from splenocytes of an H.B19 chicken were isolated by flow cytometry and expanded in the presence of chicken interleukin (IL)-2 to produce a primary NK cell line; other lines are standard chicken cell lines (T.G., unpublished data). Standard RNA preparation and polymerase chain reaction with reverse transcriptase (RT-PCR) using the primers 5'-AACTTCTGAGCCATGGATGAG-3' and 5'-TGGATGGAGACGTAAAGGTTTC-3' for B-NK1 and primers 5'-TACCACAATGTACCCCTGGC-3' and 5'-CTCGTCTGTGTTTATGCGC-3' for actin generated fragments, which were cloned and sequenced (Fig. 2a) or analysed by gel electrophoresis and ethidium bromide staining (Fig. 2b).

Received 21 July; accepted 27 August 1999.

1. Briles, W., McGibbon, W. & Irwin, M. On multiple alleles affecting cellular antigens in the chicken. *Genetics* **35**, 633–640 (1950).
2. Pink, J., Droege, W., Hala, K., Miggiano, V. & Ziegler, A. A three-locus model for the chicken major histocompatibility complex. *Immunogenetics* **5**, 203–216 (1977).
3. Guillemot, F. *et al.* A molecular map of the chicken major histocompatibility complex: the class IIβ genes are closely-linked to the class I genes and the nucleolar organizer. *EMBO J.* **7**, 2775–2785 (1988).
4. Kaufman, J., Völk, H. & Wallny, H. J. A "minimal essential Mhc" and an "unrecognized Mhc": two extremes in selection for polymorphism. *Immunol. Rev.* **143**, 63–88 (1995).
5. MHC Sequencing Consortium. Complete sequence and gene map of a human major histocompatibility complex. *Nature* **401**, 921–923 (1999).
6. Pamer, E. & Cresswell, P. Mechanisms of MHC class I-restricted antigen processing. *Annu. Rev. Immunol.* **16**, 323–358 (1998).
7. Lanier, L. NK cell receptors. *Annu. Rev. Immunol.* **16**, 359–393 (1998).
8. Briles, W., Briles, R., Taffs, R. & Stone, H. Resistance to a malignant lymphoma in chickens is mapped to a subregion of the major histocompatibility (B) complex. *Science* **219**, 977–979 (1983).
9. Plachy, J., Pink, J. & Hala, K. Biology of the chicken MHC (B complex). *Crit. Rev. Immunol.* **12**, 47–79 (1992).
10. Kaufman, J. & Lamont, S. In *The Major Histocompatibility Complex Region in Domestic Animal Species* (eds Schook, L.B. & Lamont, S. L.) (CRC, Boca Raton, 1996).
11. Miller, M. *et al.* Assignment of *Rfp-Y* to the chicken MHC/NOR microchromosome and evidence for high frequency recombination associated with the nucleolar organizer region. *Proc. Natl Acad. Sci. USA* **93**, 3958–3962 (1994).
12. Pazderka, F., Longenecker, B., Law, G. & Ruth, R. The major histocompatibility complex of the chicken. *Immunogenetics* **2**, 101–130 (1975).
13. Vainio, O., Veromaa, T., Eerola, E., Toivanen, P. & Ratcliffe, M. Antigen-presenting cell-T cell interaction in the chicken is MHC class II antigen restricted. *J. Immunol.* **140**, 2864–2868 (1988).
14. Vainio, O., Koch, C. & Toivanen, A. B-L antigens (class II) of the chicken major histocompatibility complex control T-B cell interaction. *Immunogenetics* **19**, 131–140 (1984).
15. Thorpe, K., Abdulla, S., Kaufman, J., Trowsdale, J. & Beck, S. Phylogeny and structure of the RING3 gene. *Immunogenetics* **44**, 391–396 (1996).
16. Kaufman, J., Bustead, N., Miller, M., Riegert, P. & Salomonsen, J. In *Advances in Avian Immunology Research* (eds Davison, T. E., Bumstead, N. & Kaiser, P.) 119–127 (Carfax Publishing, Abingdon, 1995).
17. Frangoulis, B. *et al.* Identification of the Tapasin gene in the chicken major histocompatibility complex. *Immunogenetics* **49**, 328–337 (1999).
18. Nonaka, M. *et al.* Major histocompatibility complex gene mapping in the amphibian *Xenopus* implies a primordial organisation. *Proc. Natl Acad. Sci. USA* **94**, 5789–5791 (1997).
19. Kaufman, J., Skjødt, K. & Salomonsen, J. The B-G multigene family of the chicken MHC. *Crit. Rev. Immunol.* **11**, 113–143 (1991).
20. Henry, J., Miller, M. & Pontarotti, P. Structure and evolution of the extended B7 family. *Immunol. Today* **20**, 285–288 (1999).
21. Kandil, E. *et al.* Isolation of low molecular mass polypeptide complementary clones from primitive vertebrates. *J. Immunol.* **156**, 4245–4253 (1996).
22. Rammensee, H.-G., Friede, T. & Stevanovic, S. MHC ligands and peptide motifs: first listing. *Immunogenetics* **41**, 178–228 (1995).
23. Bumstead, N. Genomic mapping of resistance to Marek's disease. *Avian Pathol.* **27**, s78–s81 (1998).
24. Skjødt, K., Koch, C., Crone, M. & Simonsen, M. Analysis of chickens for recombination with the MHC (B complex). *Tissue Antigens* **25**, 278–282 (1985).
25. Germain, R., Bentley, D. & Quill, H. Influence of allelic polymorphism on the assembly and surface expression of class II MHC (Ia) molecules. *Cell* **43**, 233–242 (1985).
26. Momburg, F. *et al.* Selectivity of MHC-encoded peptide transporters from human, mouse and rat. *Nature* **367**, 648–651 (1994).
27. Joly, E. *et al.* Co-evolution of rat TAP transporters and MHC class I RT1-A molecules. *Curr. Biol.* **8**, 169–172 (1998).
28. Pharr, G., Dodgson, J., Hunt, H. & Bacon, L. Class II MHC cDNAs in 12L₁ B-congenic chickens. *Immunogenetics* **5**, 350–354 (1998).
29. Bacon, L., Ismail, N. & Motta, J. Allograft and antibody responses of 15L₁ B congenic chickens. *Prog. Clin. Biol. Res.* **238**, 219–233 (1987).

Acknowledgements

We thank I. Shaw, T. Powell, S. Rogers, A. Smith, P. Kaiser and G. Griffiths for help with analysis and/or critical review of the manuscript. S.M. and S.B. are funded by the Wellcome Trust and, at the early stage of the project, by the Imperial Cancer Research Fund and a grant from the EU BioMed 1 programme. J.J. is supported by the EU FAIR program.

Correspondence and requests for materials should be addressed to J.K.
(e-mail: jim.kaufman@bbsrc.ac.uk).

Actin-based motility of vaccinia virus mimics receptor tyrosine kinase signalling

Friedrich Frischknecht, Violaine Moreau, Sabine Röttger, Stefania Gonfloni*, Inge Reckmann, Giulio Superti-Furga* & Michael Way

Cell and * Developmental Biology Programmes, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany

Studies of the actin-based motility of the intracellular pathogens *Listeria monocytogenes* and *Shigella flexneri* have provided important insight into the events occurring at the leading edges of motile cells^{1–5}. Like the bacteria *Listeria* and *Shigella*, vaccinia virus, a relative of the causative agent of smallpox, uses actin-based motility to spread between cells⁶. In contrast to *Listeria* or *Shigella*, the actin-based motility of vaccinia is dependent on an unknown phosphotyrosine protein, but the underlying mechanism remains obscure⁷. Here we show that phosphorylation of tyrosine 112 in the viral protein A36R by Src-family kinases is essential for the actin-based motility of vaccinia. Tyrosine phosphorylation of A36R results in a direct interaction with the adaptor protein Nck⁸ and the recruitment of the Ena/VASP family member N-WASP⁹ to the site of actin assembly. We also show that Nck and N-WASP are essential for the actin-based motility of vaccinia virus. We suggest that vaccinia virus spreads by mimicking the signalling pathways that are normally involved in actin polymerization at the plasma membrane.

We began our investigation into the mechanism of actin-based motility of vaccinia virus by identifying the phosphotyrosine protein observed at the site of actin assembly⁷. Western-blot analysis reveals that three prominent proteins, pTyr200, pTyr80/85 and pTyr50, consistently become tyrosine phosphorylated during vaccinia infection (Fig. 1). Western-blot analysis and immunoprecipitation experiments identified pTyr80/85 as the actin-binding protein cortactin¹⁰ (data not shown), which is known to be enriched in the cell cortex and becomes phosphorylated by c-Src when *Shigella* enters the cell¹¹. To determine whether tyrosine phos-

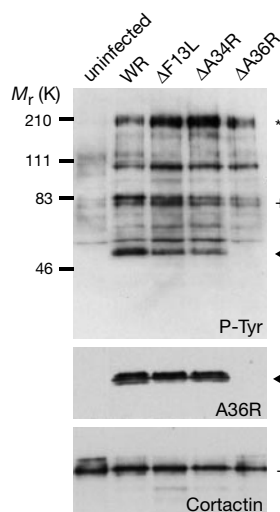


Figure 1 Vaccinia A36R protein is tyrosine phosphorylated. Top, western-blot analysis of phosphotyrosine proteins during vaccinia infection reveals that three proteins, pTyr200 (asterisk), pTyr80/85 (cross) and pTyr50 (arrowhead) consistently become tyrosine phosphorylated, although pTyr50 (arrowhead) is absent in Δ A36R infections. Bottom, the same blot probed with A36R and cortactin antibodies. Virus strains are indicated.

phorylation of pTyr200, pTyr80/85 and pTyr50 was dependent on actin tail formation, we examined the phosphorylation pattern of cells infected with the recombinant viral strains Δ A34R, Δ A36R and Δ F13L, which do not induce actin tails at 8 hours post-infection^{12,13}. We found that pTyr50 was absent in cells infected with a recombinant virus lacking A36R (Fig. 1). Western-blot analysis of two-dimensional gels and immunoprecipitation experiments showed that pTyr50 and A36R are the same protein (data not shown). The vaccinia A36R protein is a type Ib integral membrane protein with a cytoplasmic domain of about 200 amino-acid residues¹³ that plays

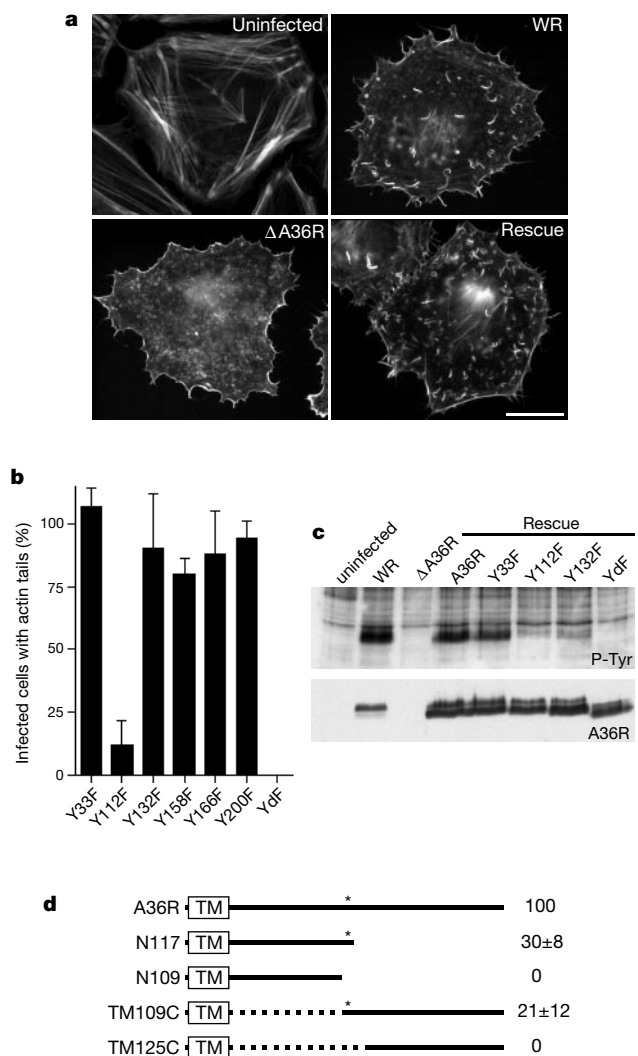


Figure 2 Tyrosine phosphorylation of A36R is essential for vaccinia actin tail formation. **a**, Immunofluorescence analysis reveals that ectopic expression of A36R in the Δ A36R background rescues actin tail formation. Scalebar, 20 μ m. **b**, Quantification of actin tail formation by all six A36R tyrosine-to-phenylalanine mutants reveals that only Y112 is required for efficient actin tail formation. However, only the A36R–YdF double mutant (Y112F and Y132F combined) eliminates all actin tail formation. **c**, Top, western-blot analysis of ectopically expressed A36R point mutants reveals that Δ A36R-infected cells expressing Y112F or Y132F have a weaker pTyr50 signal than cells infected with WR, or with Δ A36R expressing wild-type A36R or Y33F. Y158F, Y166F and Y200F mutants are phosphorylated to a similar extent as A36R or Y33F (data not shown). No pTyr50 signal is detectable in Δ A36R-infected cells expressing the YdF double mutant. Bottom, the same blot reprobed with anti-A36R antibody. YdF migrates slightly faster than the other A36R proteins, consistent with an absence of tyrosine phosphorylation. **d**, Diagram of A36R and the deletion mutants N109, N117, TM109C and TM125C (see Methods for nomenclature). The box at the N terminus shows the predicted transmembrane domain (TM). Asterisk indicates Y112; dotted line indicates the region of A36R that is deleted in TM109C and TM125C; numbers on the right show the rescue efficiencies of actin tail formation.

an essential but undefined role in the actin-based motility of vaccinia^{12,14}.

To study the role of A36R phosphorylation in vaccinia actin-based motility, we tested whether ectopic expression of the wild-type A36R protein in Δ A36R infected cells would rescue actin tail formation. Ectopic expression of A36R resulted in both tyrosine phosphorylation of the protein and efficient rescue of actin tail formation ($111 \pm 36\%$) (Fig. 2a, c). Expression of point mutants of the six individual tyrosine residues in the A36R cytoplasmic domain showed that only a change of tyrosine 112 to phenylalanine (A36R mutant Y112F) resulted in a dramatic reduction of actin tails (Fig. 2b). However, western-blot analysis revealed that mutation of either tyrosine 112 or tyrosine 132 reduced A36R phosphorylation, indicating that both of these residues are phosphorylated (Fig. 2c). As A36R–Y112F still rescues actin tail formation, albeit at low levels, we produced the double point mutant, YdF, in which both residues 112 and 132 were changed to phenylalanine. Although expression of A36R–YdF was similar to the other point mutants, we found no evidence for either tyrosine phosphorylation of the protein or rescue of actin tail formation (Fig. 2b, c). Lack of rescue was not due to mistargeting, as A36R–YdF localized to the intracellular enveloped form of vaccinia virus, which is responsible for actin tail formation⁶ (data not shown). In parallel with studies using A36R point mutants, we examined the ability of several A36R deletion constructs to rescue actin tail formation (Fig. 2d). The only requirement for actin tail formation was the presence of tyrosine 112, which represents a good Src-family kinase phosphorylation site¹⁵. Vaccinia infection also induces phosphorylation of cortactin, a well known Src substrate¹⁰, so we tested whether this family of kinases phosphorylates A36R. We found that the Src-family kinase-inhibitor PP1 (ref. 16) reduced tyrosine phosphorylation of both A36R and cortactin, as well as reducing actin tail formation (Fig. 3a; and data not shown). However, PP1 did not affect viral assembly or A36R localization (data not shown). In addition, expression of c-Src kinase 'dead open' mutants 527 Kin[−] and KP Kin[−] reduced tyrosine phosphorylation of A36R and actin tail formation (Fig. 3a, b; and data not shown). When expressed at low levels, 527 Kin[−] and KP Kin[−] were observed at the site of vaccinia actin tail formation, which is consistent with their dominant-negative activity when overexpressed (Fig. 3c).

Our observations that vaccinia actin tail formation is dependent on Src-family kinases and tyrosine phosphorylation is reminiscent

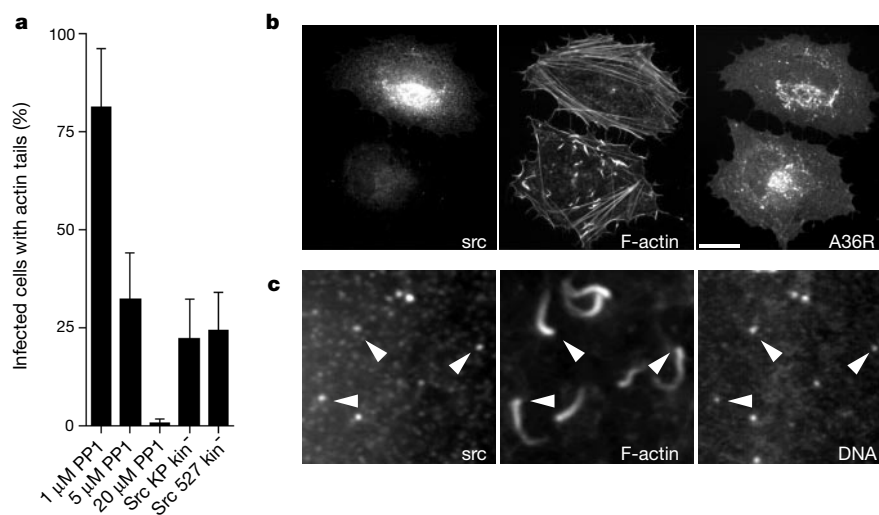


Figure 3 Src-family kinases mediate actin tail formation of vaccinia virus. **a**, The Src-family kinase inhibitor PP1 inhibits actin tail formation in a dose-dependent manner. Overexpression of dead open c-Src mutants also reduces the efficiency of actin tail formation. **b**, Immunofluorescence analysis of the effects of overexpressing dead open

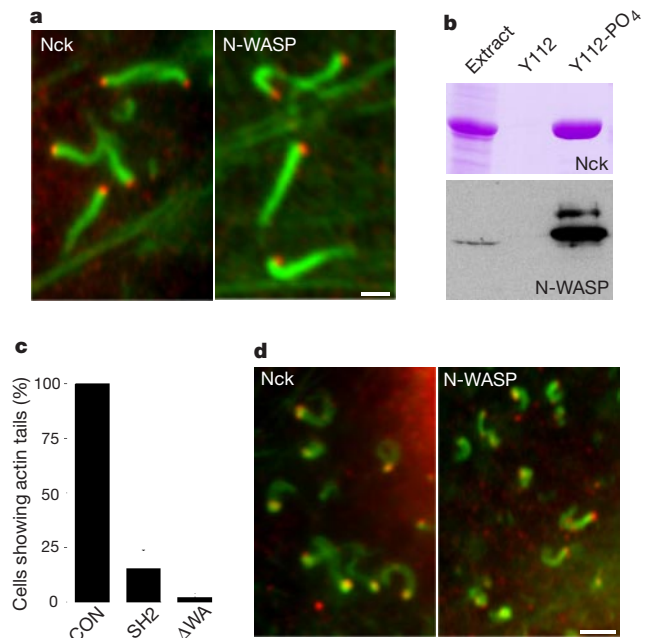


Figure 4 Nck and N-WASP are essential for vaccinia actin tail formation.

a, Immunofluorescence images showing that Nck and N-WASP (red) are recruited to vaccinia virus particles that have induced actin tails (green). Scale bar, 2 μ m. **b**, Top, Coomassie-stained gel showing that the peptide A36R105–116 containing phosphorylated tyrosine 112 interacts directly with Nck, as it retains the protein from soluble *Escherichia coli* extract. Y112 and Y112-PO₄ represent the unphosphorylated and phosphorylated versions of the A36R105–116 peptide, respectively. Bottom, western blot showing that the Nck-A36R105–116 complex recruits N-WASP from extracts prepared from infected cells. The resin sample represents five times more material than the extract sample, demonstrating there is no recruitment of N-WASP by the unphosphorylated A36R105–116 peptide. **c**, Overexpression of the Nck-SH2 domain (SH2) or an N-WASP construct lacking the WA domain (Δ WA) inhibits actin tail formation in WR-infected cells. CON, control cells infected with WR but not transfected. **d**, Immunofluorescence images showing that Nck and N-WASP (red) are recruited to clathrin-coated vesicles that can nucleate actin tails (green) in uninfected cells. Scale bar, 2 μ m.

c-Src mutants on vaccinia actin tail formation. Two infected cells, one transfected (top) and one untransfected (bottom), are shown. Scale bar, 20 μ m. **c**, Closer examination of infected cells expressing low levels of dead open c-Src reveals that the protein is recruited to viral particles that have induced actin tails (arrowheads).

of signal-transduction pathways involved in the control of actin polymerization at the plasma membrane. This analogy is strengthened by the localization of the adaptor protein Nck⁸ at the site of vaccinia actin tail formation (Fig. 4a). Nck recruitment was dependent on both A36R and its phosphorylation, as it was absent in Δ A36R infections, with or without ectopic expression of A36R–YdF (data not shown). *In vitro* binding assays using peptides corresponding to residues 105–116 of A36R demonstrate that recruitment of Nck to the virus occurs through the direct interaction with A36R and is dependent on phosphorylation of tyrosine 112 (Fig. 4b). Furthermore, overexpression of the Nck SH2 domain resulted in a dramatic reduction of actin tail formation (Fig. 4c). The involvement of Nck raises the question whether its downstream partner WASP^{8,17,18} is also required for vaccinia actin tail formation. We found that N-WASP⁹, a member of the Ena/VASP family, is recruited to the virus (Fig. 4a). N-WASP was also efficiently recruited from cell extracts by the Nck–A36R-phosphorylated peptide complex (Fig. 4b). As observed with the Nck SH2 domain, overexpression of a dominant-negative construct, N-WASP– Δ WA¹⁹, inhibits actin tail formation (Fig. 4c). The requirement of N-WASP, which is known to stimulate the actin-nucleating activity of the Arp2/3 complex^{20,21}, seems to provide the final link to achieve vaccinia actin tail formation. Nck and N-WASP, together with the Arp2/3 complex, are also observed on clathrin-coated vesicles that can nucleate little actin tails⁷ (Fig. 4), indicating that this complex of proteins plays an important role in initiating actin polymerization in the cell.

Vaccinia virus therefore achieves actin-based motility by mimicking the receptor tyrosine kinase signalling pathways that control actin polymerization at the plasma membrane. The sequence conservation between vaccinia A36R and its homologue A39R in variola virus, the causative agent of smallpox, indicates that variola virus also used actin-based motility to spread between cells. This Src-family kinase-dependent signalling pathway may also account for other cellular effects induced by vaccinia infection, including induction of cell motility, loss of contact inhibition and changes in cell adhesion, all of which are reminiscent of a transformed phenotype^{22,23}. We suggest that vaccinia virus provides an excellent model system in which to study Src-family kinase and the signalling pathways that affect actin polymerization, adhesion and cell motility. □

Methods

Western-blot and immunofluorescence analysis

Western reserve and recombinant vaccinia strains that lack the genes A34R, A36R or F13L are referred to as WR, Δ A34R, Δ A36R and Δ F13L¹². HeLa cells were infected and processed for western-blot or immunofluorescence analysis, and images were acquired as described^{7,13}. Intracellular enveloped viruses were labelled with A33R or A36R antibodies¹³, and tyrosine-phosphorylated proteins were detected with the 4G10 monoclonal antibody (Upstate Biotechnology). Src was detected by using anti-Src 327 mouse monoclonal antibody²⁴. Antibodies against cortactin and Nck were obtained from Upstate Biotechnology, Transduction Laboratories and Santa Cruz Biotechnology. Rhodamine- and Alexa 488-phalloidin and PP1 were obtained from Molecular Probes and Alexis Corporation, respectively.

Construction of pEL A36R expression vectors

A36R expression was placed under the control of the synthetic vaccinia virus early/late promoter (E/L)²⁵ by polymerase chain reaction (PCR) using the primers PELA36RFor 5'-CCGCGCTCGAGAAAATGAAATTTTATTTTATTTTATTTTGGATATATAA-TAAGCTCGAGATCTACCATGATGCTGGTACCTCTTATC-3' and A36RH3Rev 5'-GGCAAGCTTATCACCAATGATACGACCGATGATTCTAT-3'. The resulting PCR product was cloned into the *Xho*I–*Hind*III sites of *Not*I-deficient pBluescript SKII to generate the pELA36R construct. Tyrosine-to-phenylalanine point mutations of A36R were engineered by 'round the world' PCR²⁶ using pELA36R as a template. The pELA36R–YdF construct was engineered in the same way using A36R–Y112F as the template. The deletion mutants N109 and N117 were generated by introducing a stop codon after residue 109 and 117, respectively, and are named according to the remaining sequence, that is, amino-terminal to their final carboxy-terminal residue (Fig. 2d). We took advantage of a unique SnaB1 site, which immediately follows the transmembrane (TM) domain of A36R (residues 1–32) to generate internal deletions by PCR. Internal deletions TM109C and TM125C are defined by the presence of the transmembrane domain followed by the first residue of the cytoplasmic domain of A36R (109 or 125) extending to the C terminus (Fig. 2d).

pEL dominant-negative expression constructs

*Eco*RI–*Bam*HI inserts containing the c-Src mutants 527 Kin[–] and KP Kin[–] were cloned into the *Eco*RI–*Bgl*II sites of the pELA36R vector backbone. The 527 and KP of the K295M catalytically dead (Kin[–]) c-Src represent mutations of Y527F and K249G/P250E²⁴, respectively. These proteins represent 'dead open' conformations of c-Src which are able to bind cellular targets but are catalytically inactive. EGFP was amplified by PCR from the pEGFP-N1 expression vector (Clontech) and cloned into the *Bgl*III–*Xho*I sites of pEL immediately downstream of the viral promoter. The DNA coding residues 1–397 of rat N-WASP⁹ were then amplified by PCR and cloned into the *Not*I–*Eco*RI sites of the pELGFP vector to generate pELGFP N-WASP– Δ WA, which lacks the WH2 and acidic domains of the protein¹⁹. The DNA corresponding to the SH2 domain, residues 275–373, of human Nck1 and the full-length protein were amplified by PCR from a HeLa cDNA library and cloned in an identical fashion to N-WASP– Δ WA into pELGFP. The fidelity of all pEL constructs was confirmed by sequencing.

Expression and quantification of actin tail rescue or inhibition

pEL constructs were transfected into HeLa cells 4 hours post-infection with either WR or Δ A36R viral strains and processed for immunofluorescence or western-blot analysis 4 hours later. Similarly, PP1 was added to cells either before infection or up to 4 hours post-infection. Rescue of actin tail formation or inhibition efficiencies were determined from over 200 infected cells in 4 independent experiments based on criteria described previously⁷. All data were normalized to WR or A36R rescue experiments where appropriate, and error bars indicate standard deviation of the mean.

Protein expression and *in vitro* binding

Human Nck1 was cloned from pELGFP into the vector pMW172, expressed in BL21(DE3), and the soluble fraction prepared as described²⁷. The tyrosine-phosphorylated and unphosphorylated versions of the peptide CCGAPSTEHIYDSVAGST, corresponding to residues 105–116 of A36R, were coupled by the N-terminal cysteine residue to Sulfolink resin (Pierce). The BL21(DE3) soluble fraction containing Nck1 was incubated with peptide resin in the presence of 1 mM vanadate at room temperature for 30 min and washed 5 times with phosphate-buffered saline containing 100 mM NaCl, 0.1% Tween and 1 mM vanadate. The washed resin was then incubated with cell extracts prepared as described¹³. Samples were removed for SDS–PAGE and western-blot analysis to assess Nck1 and N-WASP binding.

Received 23 June; accepted 19 August 1999.

1. Dramsi, S. & Cossart, P. Intracellular pathogens and the actin cytoskeleton. *Annu. Rev. Cell Dev. Biol.* **14**, 137–166 (1998).
2. Finlay, B. B. & Cossart, P. Exploitation of mammalian host cell functions by bacterial pathogens. *Science* **276**, 718–725 (1997).
3. Welch, M. D., Iwamatsu, A. & Mitchison, T. J. Actin polymerization is induced by Arp2/3 protein complex at the surface of *Listeria monocytogenes*. *Nature* **385**, 265–269 (1997).
4. Welch, M. D., Rosenblatt, J., Skole, J., Portnoy, D. A. & Mitchison, T. J. Interaction of human Arp2/3 complex and the *Listeria monocytogenes* ActA protein in actin filament nucleation. *Science* **281**, 105–108 (1998).
5. Mullins, D. R., Heuser, J. A. & Pollard, T. D. The interaction of Arp2/3 complex with actin: nucleation, high affinity pointed end capping, and formation of branching networks of filaments. *Proc. Natl. Acad. Sci. USA* **95**, 6181–6186 (1998).
6. Cudmore, S., Cossart, P., Griffiths, G. & Way, M. Actin-based motility of vaccinia virus. *Nature* **378**, 636–638 (1995).
7. Frischknecht, F. *et al.* Tyrosine phosphorylation is required for actin based motility of vaccinia but not *Listeria* or *Shigella*. *Curr. Biol.* **9**, 89–92 (1999).
8. McCarty, J. H. The Nck SH2/SH3 adaptor protein: a regulator of multiple intracellular signal transduction events. *BioEssays* **20**, 913–921 (1998).
9. Miki, H., Minura, K. & Takenawa, T. N-WASP, a novel actin-depolymerizing protein, regulates the cortical cytoskeletal rearrangement in a PIP2-dependent manner downstream of tyrosine kinases. *EMBO J.* **15**, 5326–5335 (1996).
10. Wu, H. & Parsons, J. T. Cortactin, an 80/85-kilodalton pp60src substrate, is a filamentous actin-binding protein enriched in the cell cortex. *J. Cell Biol.* **120**, 1417–1426 (1993).
11. Dehio, C., Prevost, M. C. & Sansonetti, P. J. Invasion of epithelial cells by *Shigella flexneri* induces tyrosine phosphorylation of cortactin by a pp60c-src-mediated signalling pathway. *EMBO J.* **14**, 2471–2482 (1995).
12. Sanderson, C. M., Frischknecht, F., Way, M., Hollinshead, M. & Smith, G. L. Roles of vaccinia virus EEV-specific proteins in intracellular actin tail formation and low pH-induced cell-cell fusion. *J. Gen. Virol.* **79**, 1415–1425 (1998).
13. Röttger, S., Frischknecht, F., Reckmann, I., Smith, G. L. & Way, M. Interactions between vaccinia virus IEV membrane proteins and their roles in IEV assembly and actin tail formation. *J. Virol.* **73**, 2863–2875 (1999).
14. Wolffe, E. J., Weisberg, A. S. & Moss, B. Role for the vaccinia virus A36R outer envelope protein in the formation of virus-tipped actin-containing microvilli and cell-to-cell virus spread. *Virology* **25**, 20–26 (1998).
15. Songyang, Z. & Cantley, L. C. Recognition and specificity in protein tyrosine kinase-mediated signalling. *Trends Biochem. Sci.* **20**, 470–475 (1995).
16. Hanke, J. H. *et al.* Discovery of a novel, potent, and src family selective tyrosine kinase inhibitor. *J. Biol. Chem.* **271**, 695–701 (1996).
17. Rivero-Lezcano, O. M. & Marcilla, A., Sameshima, J. H. & Robbins, K. C. Wiskott-Aldrich syndrome protein physically associates with Nck through Src3 homology domains. *Mol. Cell Biol.* **15**, 5725–5731 (1995).
18. Snapper, S. B. & Rosen, F. S. The Wiskott-Aldrich syndrome protein (WASP): roles in signalling and cytoskeletal organization. *Annu. Rev. Immunol.* **17**, 905–929 (1999).
19. Machesky, L. M. & Insall, R. H. Scar1 and the related Wiskott-Aldrich syndrome protein WASP regulate the actin cytoskeleton through the Arp2/3 complex. *Curr. Biol.* **8**, 1347–1356 (1998).

20. Machesky, L. M. *et al.* Scar, a WASP-related protein, activates dendritic nucleation of actin filaments by the Arp2/3 complex. *Proc. Natl Acad. Sci. USA* **96**, 3739–3744 (1999).
21. Rohatgi, R. *et al.* The interaction between N-WASP and the Arp2/3 complex links Cdc42-dependent signals to actin assembly. *Cell* **97**, 221–231 (1999).
22. Sanderson, C. M., Way, M. & Smith, G. L. Virus-induced cell motility. *J. Virol.* **72**, 1235–1243 (1998).
23. Sanderson, C. M. & Smith, G. L. Vaccinia virus induces calcium-independent cell-matrix adhesion during the motile phase of infection. *J. Virol.* **72**, 9924–9933 (1998).
24. Gonfloni, S. *et al.* The role of the linker between the SH2 domain and catalytic domain in the regulation and function of Src. *EMBO J.* **16**, 7261–7271 (1997).
25. Chakrabarti, S., Sisler, J. R. & Moss, B. Compact, synthetic, vaccinia virus early/late promoter for protein expression. *Biotechniques* **23**, 1094–1097 (1997).
26. Hemsley, A., Arnheim, N., Toney, M. D., Cortopassi, G. & Galas, D. J. A simple method for site-directed mutagenesis using the polymerase chain reaction. *Nucleic Acids Res.* **17**, 6545–6551 (1989).
27. Way, M., Pope, B. & Weeds, A. G. Evidence for functional homology in the F-actin binding domain of gelsolin and alpha-actinin: implications for the requirements of severing and capping. *J. Cell Biol.* **119**, 835–842 (1992).

Acknowledgements

We are grateful to H. Miki for N-WASP cDNA and N-WASP antibody. We thank G. Smith and R. Blasco for providing recombinant vaccinia strains; J. White, C. Blaumüller, A. Desai and A. Poubidou for reading the manuscript; T. Harder for encouraging us to try PP1; and S. Guth for the 'round the world' PCR method. S.G. was supported by a fellowship from the Boncompagni-Ludovisi Foundation.

Correspondence and requests for materials should be addressed to M.W. (e-mail: way@embl-heidelberg.de).

Asynchronous replication of imprinted genes is established in the gametes and maintained during development

Itamar Simon^{*}, Toyooki Tenzen^{†*}, Benjamin E. Reubinf[‡], Dahlia Hillman[‡], John R. McCarrey[§] & Howard Cedar^{*}

^{*} Department of Cellular Biochemistry, Hebrew University, Jerusalem 91120, Israel

[†] Department of Evolutionary Genetics, National Institute of Genetics, Mishima, Shizuoka-ken 411-8540, Japan

[‡] Department of Obstetrics and Gynecology, Hadassah Ein-Kerem University Hospital, Jerusalem 91120, Israel

[§] Department of Genetics, Southwest Foundation for Biomedical Research, San Antonio, Texas 78228, USA

Genomic imprinting is characterized by allele-specific expression of multiple genes within large chromosomal domains¹ that undergo DNA replication asynchronously during S phase^{2,3}. Here we show, using both fluorescence *in situ* hybridization analysis and S-phase fractionation techniques, that differential replication timing is associated with imprinted genes in a variety of cell types, and is already present in the pre-implantation embryo soon after fertilization. This pattern is erased before meiosis in the germ line, and parent-specific replication timing is then reset in late gametogenesis in both the male and female. Thus, asynchronous replication timing is established in the gametes and maintained throughout development, indicating that it may function as a primary epigenetic marker for distinguishing between the parental alleles.

Fluorescence *in situ* hybridization (FISH) analysis on interphase nuclei can be used to assay the replication properties of individual alleles in nonsynchronously growing cells⁴. For most gene regions, one observes either two single or two double hybridization dots per nucleus, indicating that both alleles replicate at about the same time in S phase. In contrast, FISH analysis of mono-allelically expressed sequences, such as imprinted genes² or olfactory receptor genes⁵,

reveals a large number of nuclei in which one allele has already replicated while the other has not, thus indicating asynchronous replication. Interestingly, this pattern is regional and is found in a variety of cell types independent of expression^{2,3}, suggesting that it is also preserved during development.

Imprinted gene regions have a parent-of-origin specific replication pattern^{2,3}. In contrast, the olfactory receptor gene loci replicate in a random manner; in some cells the paternal allele replicates early, whereas in others the maternal allele replicates early⁵. This latter pattern is similar to that seen with the X chromosome in female cells.

Although the FISH approach provides a good measure of replication timing and a sensitive way to detect allelic asynchrony^{2,3,6,7}, we wanted to show directly that imprinted genes do replicate with an asynchronous pattern, as determined by quantitative polymerase chain reaction (PCR) analysis of pulse-labelled BrdU DNA⁸ from S-phase-sorted cell fractions⁹. As shown in Fig. 1a, the human *Igf2* gene in EBV-transformed lymphocytes replicates in a broad peak covering the middle of S phase. Use of a restriction-site polymorphism to distinguish between the two alleles, however, shows that one chromosomal copy clearly replicates before the other despite considerable overlap (Fig. 1a). In a similar experiment, we characterized the replication properties of the mouse *Igf2r* gene using an Abelson-transformed pre-B-cell subclone from Spretus/Musculus F₁ mice (Fig. 1b). Here, the replication pattern of each allele is clearly separable. As the pedigree of these cells is known, this system also allowed us to prove that it is the paternal Spretus allele that replicates early².

Unlike the FISH methodology, which detects the sharp endpoint of replication, the resolving power of cell-cycle fractionation techniques is low, probably because DNA replication may not initiate at exactly the same time within every individual cell, and because there is considerable overlap in the fluorescence-sorted S-phase fractions. This might help explain why in studies using similar techniques, but based on fewer S-phase fractions, the *Igf2/H19* locus did not appear to replicate asynchronously^{10,11}. In fact, a careful visual re-examination of the raw data from one of these studies¹⁰ reveals allelic differences very similar to those seen in our own PCR analyses. Thus, these results indicate that FISH analysis provides a reliable and accurate method for detecting asynchronous replication, and has the advantage that it can be used on small cell populations *in vivo*.

To understand the functional relationship between genomic

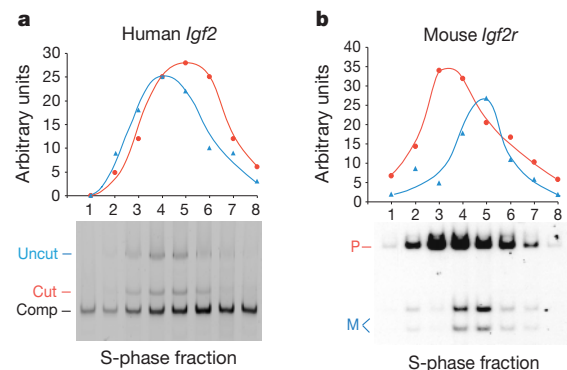


Figure 1 Detection of asynchronous timing by S-phase fractionation. Replication timing was analysed by S-phase fractionation of BrdU-labelled DNA (see Methods) using restriction enzymes to distinguish between the paternal (P) and maternal (M) alleles of *Igf2* in EBV-transformed human lymphoblasts (a) and *Igf2r* in mouse Abelson-transformed pre-B cells (b). When these same B cells were analysed by FISH, we detected 44% single/single (SS), 35% single/double (SD) and 21% double/double (DD) signals. After subtracting normal background (~18% SD), the difference between the two alleles is roughly 15%, or ~1–1.5 h, a time difference similar to that seen in the Figure. Circles, uncut alleles; triangles, cut alleles. Comp, competitor DNA product.

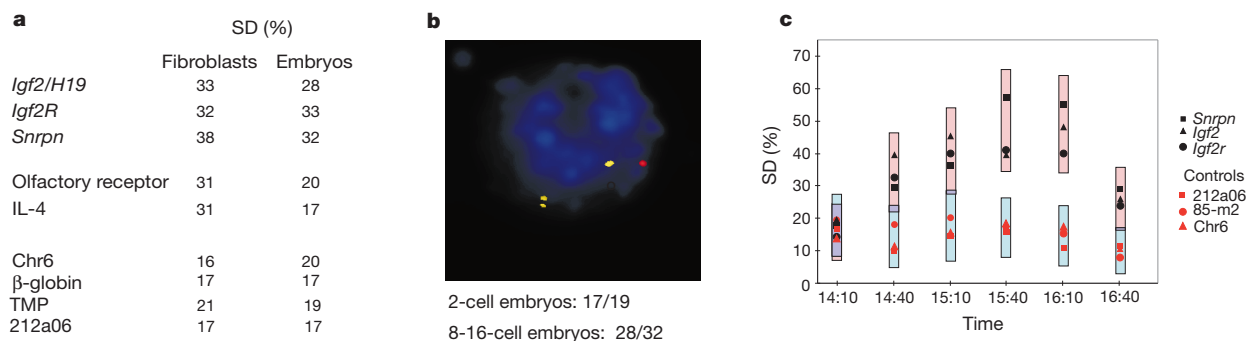


Figure 2 Replication timing in the early embryo. **a**, FISH analysis of the replication profile was carried out on primary fibroblasts or pre-implantation embryos (8–16 cells) using a variety of probes. We found a significant difference in the percentage single/double (SD) between imprinted and control groups in embryos, as well as fibroblasts ($P < 0.001$), and between the control and random groups in fibroblasts ($P = 0.002$). **b**, Nuclei from 2-cell or 8–16-cell embryos heterozygous for a chromosome 7 deletion were subject to FISH using one probe (yellow) to detect *Igf2* replication (λ up12 or 20267) and a second probe (red) for the deleted region (85-m2). These nuclei were then scored by determining the

fraction of cells with a S/D *Igf2* pattern that had the double signal associated with the paternal allele. **c**, Mixtures of zygotes from 6–8 superovulated females were incubated and assayed for SD signals by FISH at the indicated time on the day following mating using *Snrpn* (P1-clone), *Igf2r* region (LA2), *Igf2/H19* (λ up12) and control probes. Statistical variations for the three imprinted (red) or control (blue) probes are depicted by rectangles. Overlap of these limits is indicated by the striped region. Allelic analysis indicated that the paternal *Igf2* and *Igf2r* regions were early replicating in 22 out of 24 cells in which both probes had the SD configuration.

imprinting and asynchronous replication, we needed to determine how the differential pattern is set up during development. We used FISH to analyse replication timing for a number of different genes in pre-implantation 8–16-cell embryos. The imprinted *Igf2r*, *Snrpn* and *Igf2/H19* gene regions have an asynchronous pattern of replication even at this early stage of development (Fig. 2a), and experiments using mice with a defined deletion (C^{12K}) on a single copy of chromosome 7 (ref. 12) confirmed that it is almost always (90%) the paternal *Igf2* gene that replicates early. This differential pattern is clearly dependent on the presence of a maternal and paternal allele in each cell, as in parthenogenetic embryos and embryonic stem cells, *Igf2* (18% single/double (SD)), *Snrpn* (10% SD) and *Igf2r* (18% SD) all replicate synchronously.

Allele-specific replication could be seen in two-cell embryos (Fig. 2b). Furthermore, by following the appearance of SD hybridization signals in zygote pronuclei as a function of time after fertilization, we could observe asynchronous replication of imprinted genes during the first embryonic round of DNA synthesis. As shown in Fig. 2c, the number of SD dots recorded for the *Snrpn*, *Igf2r* or *Igf2/H19* gene regions goes through a peak about 16 h after fertilization. The number of SD signals observed for several control probes remained at background levels during the entire course of the experiment in these same zygotes, indicating that this asynchrony is specific to imprinted regions.

These findings are consistent with the differential replication pattern being present immediately after fertilization, and indicate that each allele may acquire its unique replication-timing properties during gametogenesis. To follow this process *in vivo*, we isolated pure populations of germ cells from different stages of development and analysed them using FISH. All of the imprinted sequences tested were asynchronously replicating in primordial germ cells (12.5 days postcoitum (d.p.c.)) in males and females (Fig. 3). This same differential pattern remains fixed throughout early spermatogenesis and can still be seen in primitive type-A spermatogonial stem cells isolated from 6-day-old mice. However, several imprinted gene loci (*Igf2* and *Snrpn*) undergo erasure and become synchronously replicating, as type-A and type-B spermatogonia isolated from 8-day-old mice undergo differentiation before the initiation of meiosis.

A similar developmental pattern was observed during oogenesis, in which both the *Igf2r* and *Snrpn* genes became synchronously replicating at 13.5 d.p.c. Although the erasure of asynchrony occurs at markedly different developmental stages in males and females, it appears to be correlated with the onset of meiosis in both cases¹³.

The decision to place a parental mark on the paternal *Igf2r* and maternal *Igf2/H19* gene copies through the establishment of a new synchronous replication time is probably also initiated during gametogenesis, as these regions already replicate in an allele-specific manner at the next round of replication in the zygote (Fig. 2c).

Strikingly, erasure of differential replication appears to be accompanied by a resetting of the clock in a gamete-specific manner. The *Igf2* locus, for example, becomes early replicating before the beginning of meiosis in the male gonad, whereas the *Igf2r* gene undergoes

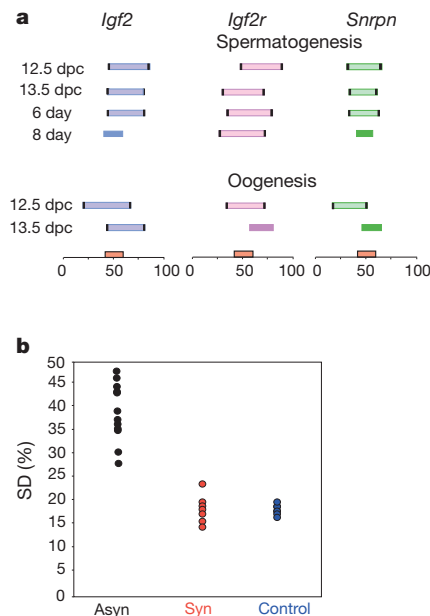


Figure 3 Replication timing during gametogenesis. **a**, FISH analysis of the replication profiles was carried out on gamete-derived cells; the single/double fraction in its proper position in S phase is shown on a scale of 0–100%. For example, the *Igf2* gene in oogenesis at 12.5 d.p.c. showed 23% single/single (SS), 47% single/double (SD) and 30% double/double (DD). All data were normalized to the pattern of a control gene from chromosome 6 (red). Probes: *Igf2*, λ up12; *Igf2r*, LA2; *Snrpn*, P1-clone. **b**, The percentage SD for each experiment was plotted according to asynchronous (Asyn), synchronous (Syn) and control groups. Two control probes (85-m2 and 212a06) and alternate probes for the *Snrpn* (plasmid P3-BamH1) and *Igf2* (cosIGF-5) regions are included. The synchronous group came out significantly different from the asynchronous group, but not different from the control group at a level of significance of 5% using a multiple comparison procedure³⁰.

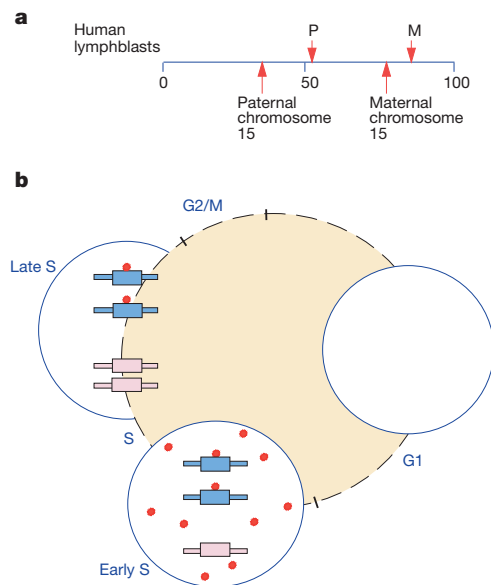


Figure 4 Maintenance of replication timing. **a**, The replication profile of the paternal (P) and maternal (M) *SNRPN* (Oncor) alleles in a human lymphoblast cell line is marked by the upper arrows on the S-phase axis (0–100%). Replication timing (% single/single (SS)) for the same *SNRPN* probe was also determined (lower arrows) in the mouse T75.2maz-34-4a cell line which contains a single maternal human chromosome 15 and the A9 + 15 line which contains a single paternal human chromosome 15 (ref. 16). The Chr6 control probe yielded similar replication-timing patterns in both cell lines (~43% SS). **b**, We propose that replication timing is controlled by *trans*-acting factors which interact with *cis*-acting elements, at the time of replication. Early replication, for example, may be triggered by an early S-phase-specific factor (red) bound to a replication-timing control element (box). Each sister chromatid of the marked allele (blue) will bind the same factor again at the time of replication, thereby providing a new mark that will ultimately direct early replication in the next cell cycle. The allele that initially lacks this factor (pink), will not be replicated in early S phase; moreover, when it finally does undergo replication later in S phase, the *trans*-acting replication factor is no longer present in the nucleus and thus cannot bind to the recognition site on this allele.

a shift to late replication during the process of oogenesis (Fig. 3). In both gametes, the *Snrpn* gene region adopts a synchronous replication pattern which is centred at about the same time zone in S phase but, at this point in gametogenesis, has apparently not yet shifted to its parent of origin-specific replication time slot, an event which must take place before the first round of replication in the zygote (Fig. 2c). This indicates that erasure and resetting of the clock may be independent processes which do not necessarily take place simultaneously. Once established, the newly set replication time for each parental allele is apparently preserved after fertilization and then maintained in each cell generation throughout development.

These molecular events are unique to the imprinted genes. Both olfactory receptor loci⁵ and the mono-allelically expressed interleukin (IL)-4 gene^{14,15} undergo asynchronous replication in somatic cells (Fig. 2a); however, this pattern is erased in the early embryo, in which these regions become synchronously replicating (Fig. 2a). Unlike imprinted genes, the differential replication pattern of these sequences is established later in development, probably at about the time of implantation, and by then there is no memory of parental origin.

As each imprinted allele can replicate at different times in S phase within a single nucleus, there must be a *cis*-acting molecular mechanism for maintaining replication timing. To show this experimentally, we examined individual copies of human chromosome 15 inserted into mouse somatic cells. This foreign DNA not only retains its imprinted expression pattern¹⁶, but also preserves its unique parent-specific time of replication at the PWS/AS locus. The paternal allele autonomously replicates at an early time in S phase,

whereas the maternally derived chromosome replicates late (Fig. 4a). These results imply the existence of a mechanism for maintaining alternate epigenetic states throughout cell division. In Fig. 4b, we propose that replication timing may be controlled by cell-cycle-dependent *trans*-acting factors that bind to recognition sites on the DNA exclusively at the time they undergo replication. Once the initial allele-specific binding is set up in the gametes, differential replication time will be preserved automatically by virtue of the controlled appearance of these key factors at the right time in the cell cycle.

Correlative evidence indicates that DNA methylation may provide a means for setting up and maintaining allelic differences¹⁷, but it is unlikely that this is the only mechanism of imprinting. Indeed, some methyl groups at key imprinting sites are transiently erased from the genome during pre-implantation development¹⁸, and imprints can be formed on DNA fragments injected into gamete-specific pronuclei even though the actual methylation is delayed for several cell divisions¹⁹. Also, differential methylation is not always correlated with mono-allelic gene expression^{20,21}, and some genes can retain their imprinted status in DNA-methyltransferase-deficient embryos²². Such findings indicate that there must be other underlying epigenetic markers that may cooperate with methylation to preserve allelic differences during development.

We have shown that asynchronous replication timing is established in the gametes and maintained in the embryo, making it a suitable marker during embryogenesis which is also well correlated with imprinted expression patterns. Genetic evidence indicates that both the human PWS/AS region on chromosome 15 and the mouse *Igf2/H19* imprinted domain have central control elements that are involved in coordinating imprinted gene expression and structure over very large genomic distances^{23,24}, and in some cases where these sequences are deleted, the asynchronous replication-timing pattern is completely disrupted^{7,25}. Our data indicate that replication-timing control is an integral part of the imprinting process and lay the groundwork for understanding its molecular mechanism. □

Methods

Embryos

Superovulated females were mated with (C57Black × BALB/c) F₁ or C¹² males. Zygotes were removed from the oviducts about 20–24 h after hormone injection and incubated at 37 °C in M16 medium (Sigma) to obtain pre-implantation embryos at different stages of development²⁶. Fibroblast cultures were prepared from 12.5 d.p.c. embryos²⁶. Diploid parthenogenetic embryos were made by ethanol activation and treated with cytochalasin B (5 µg ml⁻¹ for 5 h) as described²⁶.

Germ cells

Late-primordial germ cells from both male and female fetal gonads of CD-1 mice (Charles River) at 12.5 d.p.c., oogonia from fetal ovaries at 13.5 d.p.c. and type-M prospermatogonia from fetal testes at 13.5 d.p.c. were isolated by antibody-directed flow sorting as described²⁷. Primitive type-A and type-A + type-B spermatogonia were isolated from testes of 6- and 8-day-old mice by gradient sedimentation²⁸. The purity of germ cells in the recovered populations was consistently more than 85%.

FISH

All cells used for FISH analysis, including those taken from the gonads or from embryos, were first incubated in culture with 3 × 10⁻⁵ M BrdU for 1 h before harvesting to provide a marker for identifying cells in S phase. Postmeiotic germ cells do not replicate their DNA and were not amenable to this assay. Pre-implantation embryos were collected and fixed²⁹ to a poly-L-lysine slide (Sigma) by putting them into a drop of 0.01 N HCl, 0.1% Tween 20. In the case of zygotes, cells were placed on the slide at a density that allowed pronuclei pairs to be seen without interference from other nearby zygotes. After drying, the slide was washed, dehydrated, permeabilized and fixed²⁹. All other cells were treated with hypotonic KCl solution, fixed in methanol:acetic acid (3:1) and dropped on slides as described⁴. FISH detection and amplification, where necessary, were performed as described^{4,5}. Replication-timing profiles (percentage chromosomes in single/single state (SS), single/double state (SD), double/double state (DD)) were determined by counting a minimum of 100 BrdU-positive nuclei² for each cell type. For statistical analysis, the probes in each group (Fig. 2a) were treated as a single homogenous collection. Significance was calculated using a normal approximation to the binomial distribution after adjusting for multiple comparisons. The experiment involving zygote analysis (Fig. 2c) was repeated over 15 times, obtaining 15–20 zygotes per female, thus enabling us to accumulate an average of 45 nuclei for each point on the graph. Standard deviations were calculated assuming a binomial distribution.

The mouse probes were donated by: R. Shemer (P1-clone covering 100 kb of the *Snrpn* region and the plasmid P3-BamH1), S. Tilghman (plasmid 20267, 3' to *Igf2*), W. Reik (Lup12, a phage representing sequences upstream of *Igf2*), D. Barlow (pTCP and *Igf2r* gene clones LA2 and COS940), D. Ward (Chr6, a random cosmid clone from chromosome 6), R. Axel (olfactory receptor gene region phage clones 126 and 129 and YAC clone Y12), B. C. Holdener (85-M2, a BAC clone covering the deletion in C^{12x} mice), P. Fraser (mouse β -globin gene region plasmids p β 12g and β major), N. Benvenisti (15 kb TMP gene plasmid), P. Rotwein (cosIGF-5) and A. Chess (BAC clone 154f05 containing mouse IL-4 sequences, release I, Genome System Inc.). BAC clone 212a06 was randomly picked from the same library.

Replication timing by S-phase fractionation

EBV-transformed lymphoblasts, or an Abelson-transformed pre-B-cell subclone from Spretus/Musculus F₁ mice (donated by A. Chess), were labelled in 75 μ M BrdU for 45 min before harvesting, and nuclei were sorted for cell-cycle fractions according to DNA content⁹. BrdU DNA was isolated from each fraction and assayed for specific sequence content by quantitative PCR¹⁰ carried out for 29 (*Igf2*) or 25 (*Igf2r*) cycles (95 °C 40 s, 55 °C 40 s, 72 °C 40 s) using the primer pairs 5'-CTTGACCTTGAGTCAAATTGG-3' and 5'-GGTCGTGCCAATTACATTTC-3' (for human *Igf2*); and 5'-TGAGCAGTGGGCGACC TAGT-3' and 5'-CACGCGTTAGAGGATCCGCA-3' (for mouse *Igf2r*).

PCR products from the EBV DNA S-phase fractions were electrophoresed after cutting with the enzyme *Apa*I, which detects a polymorphism between the uncut (295-bp) and cut (231-bp) alleles in these human cells. Competitor DNA⁸ which has an 85-bp deletion covering the *Apa*I site, is included in every reaction mix. Following SYBR green I staining and gel scanning, the relative amount of each allele in the different fractions was normalized to the level of competitor PCR product. A similar analysis was carried out on BrdU fractions from the Abelson pre-B cell line, using the enzyme *Hae*III to distinguish between the uncut paternal Spretus and cut maternal Musculus alleles which yield 90- and 100-bp bands after digestion. In this case, PCR was carried out without added competitor, but in the presence of [α -³²P]dCTP, and detection was by autoradiography. Total uncut PCR product was first quantitated on an initial gel. Each individual band was extracted, and the proportion of maternal or paternal alleles was determined from a second gel after cutting out the PCR product.

Received 26 May; accepted 9 September 1999.

- Razin, A. & Cedar, H. DNA methylation and genomic imprinting. *Cell* **77**, 473–476 (1994).
- Kitsberg, D. *et al.* Allele-specific replication timing of imprinted gene regions. *Nature* **364**, 459–463 (1993).
- Knoll, J. H. M., Cheng, S.-D. & Lalande, M. Allele specificity of DNA replication timing in the Angelman/Prader-Willi syndrome imprinted chromosomal region. *Nature Genet.* **6**, 41–46 (1994).
- Selig, S., Okumura, K., Ward, D. C. & Cedar, H. Delineation of DNA replication time zones by fluorescence *in situ* hybridization. *EMBO J.* **11**, 1217–1225 (1992).
- Chess, A., Simon, L., Cedar, H. & Axel, R. Allelic inactivation regulates olfactory receptor gene expression. *Cell* **78**, 823–834 (1994).
- LaSalle, J. M. & Lalande, M. Domain organization of allele-specific replication within the GABRB3 gene cluster requires a biparental 15q11-13 contribution. *Nature Genet.* **9**, 386–394 (1995).
- Greally, J. M. *et al.* The mouse H19 locus mediates a transition between imprinted and non-imprinted DNA replication patterns. *Hum. Mol. Genet.* **7**, 91–95 (1998).
- Tenzen, T. *et al.* Precise switching of DNA replication timing in the GC content transition area in the human major histocompatibility complex. *Mol. Cell. Biol.* **17**, 4043–4050 (1997).
- Hansen, R. S., Canfield, T. K., Lamb, M. M., Gartler, S. M. & Laird, C. D. Association of fragile X syndrome with delayed replication of the FMR1 gene. *Cell* **73**, 1403–1409 (1993).
- Kawame, H., Gartler, S. M. & Hansen, R. S. Allele-specific replication timing in imprinted domains: absence of asynchrony at several loci. *Hum. Mol. Genet.* **4**, 2287–2293 (1995).
- Windham, L. Q. & Jones, P. A. Expression of H19 does not influence the timing of replication of the *Igf2/H19* imprinted region. *Dev. Genet.* **20**, 29–35 (1997).
- Wines, M. E., Tiffany, A. M. & Holdener, B. C. Physical localization of the mouse aryl hydrocarbon receptor nuclear translocator-2 (*Arnt2*) gene within the c112K deletion. *Genomics* **51**, 223–232 (1998).
- McCarrey, J. R. in *Cellular and Molecular Biology of the Testis* 58–89 (Oxford Univ. Press, Oxford, 1993).
- Bix, M. & Locksley, R. M. Independent and epigenetic regulation of the interleukin-4 alleles in CD4⁺ T Cells. *Science* **281**, 1352–1354 (1998).
- Riviere, I., Sunshine, M. J. & Littman, D. R. Regulation of IL-4 expression by activation of individual alleles. *Immunity* **9**, 217–228 (1998).
- Gabriel, J. M. *et al.* A model system to study genomic imprinting of human genes. *Proc. Natl Acad. Sci. USA* **95**, 14857–14862 (1998).
- Shemer, R. & Razin, A. in *Epigenetic Mechanisms of Gene Regulation* (eds Russo, V. E. A., Martienssen, R. A. & Riggs, A. D.) 215–229 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1996).
- Shemer, R. *et al.* Dynamic methylation adjustment and counting as part of imprinting mechanisms. *Proc. Natl Acad. Sci. USA* **93**, 6371–6376 (1996).
- Birger, Y., Shemer, R., Perk, J. & Razin, A. The imprinting box of the mouse *Igf2r* gene. *Nature* **397**, 84–88 (1999).
- Szabo, P. & Mann, J. R. Biallelic expression of imprinted genes in the mouse germ line: implications for erasure, establishment and mechanisms of genomic imprinting. *Genes Dev.* **9**, 1857–1868 (1995).
- Davis, T. L., Trasler, J. M., Moss, S. B., Yang, Bartolomei, M. S. Acquisition of the H19 methylation imprint occurs differentially on the parental alleles during spermatogenesis. *Genomics* **58**, 18–28 (1999).
- Caspari, T., Cleary, M. A., Baker, C. C., Guan, X.-J. & Tilghman, S. M. Multiple mechanisms regulate imprinting of the mouse distal chromosome 7 gene cluster. *Mol. Cell. Biol.* **18**, 3466–3474 (1998).
- Sutcliffe, J. S. *et al.* Deletions of a differentially methylated CpG island at the SNRPN gene define a putative imprinting control region. *Nature Genet.* **8**, 52–58 (1994).
- Leighton, P. A., Ingram, R. S., Eggenschwiler, J., Efstratiadis, A. & Tilghman, S. M. Disruption of imprinting caused by deletion of the H19 gene region in mice. *Nature* **375**, 34–39 (1995).

- Gunaratne, P. H., Nakao, M., Ledbetter, D. H., Sutcliffe, J. S. & Chinault, A. C. Tissue-specific and allele-specific replication timing control in the imprinted human Prader-Willi syndrome region. *Genes Dev.* **9**, 808–820 (1995).
- Hogan, B., Beddington, R., Costantini, F. & Lacy, E. *Manipulating the Mouse Embryo. A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, 1994).
- McCarrey, J. R., Hsu, K. C., Eddy, E. M., Klevecz, R. R. & Bolen, J. L. Isolation of viable mouse primordial germ cells by antibody-directed flow sorting. *J. Exp. Zool.* **242**, 107–111 (1987).
- Bellve, A. *et al.* Spermatogenic cells of the prepubertal mouse, isolation and morphological characterization. *J. Cell Biol.* **74**, 68–85 (1977).
- Harper, J. C. *et al.* Identification of the sex of human preimplantation embryos in two hours using an improved spreading method and fluorescent *in situ* hybridisation using directly labelled probes. *Hum. Reprod.* **9**, 721–724 (1994).
- Dunnett, C. W. A multiple comparison procedure for comparing several treatments with a control. *J. Am. Stat. Assoc.* **50**, 1096–1121 (1955).

Acknowledgements

We thank T. Jakubowicz and E. Rand for their help in preparing the manuscript and figures, B. C. Holdener for providing the C^{12x} mice and an appropriate probe for detecting this deletion, R. Nicholls for providing hybrid cell lines and W. Reik for providing the parthenogenetic ES cell line. We also thank H. Friedlander-Klar for help with the statistical analysis. This work was supported by grants from the US–Israel Binational Science foundation (H.C. and J.M.) the Israel Academy of sciences (H.C.) and the Israel Cancer Research Fund (H.C.).

Correspondence and requests for material should be addressed to H.C. (e-mail: cedar@md2.huji.ac.il).

A kinetic proofreading mechanism for disentanglement of DNA by topoisomerases

Jie Yan*, Marcelo O. Magnasco† & John F. Marko*

* Department of Physics, The University of Illinois at Chicago, 845 West Taylor Street, Chicago, Illinois 60607, USA

† Center for Studies in Physics and Biology, The Rockefeller University, 1230 York Avenue, New York, New York 10021, USA

Cells must remove all entanglements between their replicated chromosomal DNAs to segregate them during cell division. Entanglement removal is done by ATP-driven enzymes that pass DNA strands through one another, called type II topoisomerases. *In vitro*, some type II topoisomerases can reduce entanglements much more than expected, given the assumption that they pass DNA segments through one another in a random way¹. These type II topoisomerases (of less than 10 nm in diameter) thus use ATP hydrolysis to sense and remove entanglements spread along flexible DNA strands of up to 3,000 nm long. Here we propose a mechanism for this, based on the higher rate of collisions along entangled DNA strands, relative to collision rates on disentangled DNA strands. We show theoretically that if a type II topoisomerase requires an initial ‘activating’ collision before a second strand-passing collision, the probability of entanglement may be reduced to experimentally observed levels. This proposed two-collision reaction is similar to ‘kinetic proofreading’ models of molecular recognition^{2,3}.

We consider the knotting state of a single circular DNA strand. Given a ‘dumb’ topoisomerase that merely passes DNA through itself with some fixed probability each time the molecule strikes itself (Fig. 1), the knot state of a circular DNA strand will come to thermodynamical equilibrium. It will sometimes be knotted (a fraction $P_{\text{knot}}^{\text{eq}}$ of the time), and sometimes be unknotted (a fraction $P_{\text{unknot}}^{\text{eq}} = 1 - P_{\text{knot}}^{\text{eq}}$ of the time)^{4,5}. The ratio $P_{\text{knot}}^{\text{eq}}/P_{\text{unknot}}^{\text{eq}}$ is equal to the ratio of the rate at which unknotted strands are knotted to the rate at which knotted strands are unknotted.

The equilibrium knot probability $P_{\text{knot}}^{\text{eq}}$ expected theoretically for a ‘dumb’ topoisomerase^{4,5} matches results from DNA knotting

experiments allowed to reach thermal equilibrium^{6,7}. For DNA, $P_{\text{knot}}^{\text{eq}}$ depends on molecule length and ionic conditions⁴⁻⁷. Under the conditions of ref. 1, $P_{\text{knot}}^{\text{eq}} = 0.031$ for 10-kb P4 DNA and $P_{\text{knot}}^{\text{eq}} = 0.017$ for 7-kb PAB4 DNA. The more complicated problem of interlinkage of two DNA strands has also been studied in this way; the linking probability $P_{\text{link}}^{\text{eq}}$ depends on plasmid lengths and concentrations. For the conditions in ref. 1, $P_{\text{link}}^{\text{eq}} = 0.064$ for two 10-kb P4 DNA strands.

Experimental data¹ (Fig. 2) rule out the possibility that type II topoisomerases make 'random' strand passages. Certain type II topoisomerases (*Drosophila* topoisomerase II, *Escherichia coli* topoisomerase IV) suppressed the knotting of 10-kb P4 DNA to $P_{\text{knot}} = 0.00062$, and the knotting of 7-kb PAB4 DNA to $P_{\text{knot}} = 0.00019$. The mutual linking probability of P4 DNA strands was similarly suppressed to $P_{\text{link}} = 0.004$. Because these type II topoisomerases hydrolyse ATP, their strong suppression of entanglements does not violate the second law of thermodynamics. However, how these small topoisomerases ascertain the topology of large, flexible DNAs is uncertain.

Two mechanisms have been proposed: Rybenkov *et al.*¹ proposed a tracking scheme involving a topoisomerase-mediated synapse between three DNA segments, but without any quantitative analysis. Vologodskii⁸ proposed that type II topoisomerases recognize knots by their tendency to have a DNA segment inside the bend of a second DNA segment; computer simulations showed that strand passages directed as in Fig. 3 suppressed P_{knot} to as low as $0.1P_{\text{knot}}^{\text{eq}}$. Thermal equilibrium is not reached because transitions that are the reverse of that of Fig. 3 are prohibited. This model is described by Fig. 1b, but with nonequilibrium steady-state rates, and a nonequilibrium knotting probability $P_{\text{knot}} < P_{\text{knot}}^{\text{eq}}$. Thus, a nonequilibrium $P_{\text{knot}} < P_{\text{knot}}^{\text{eq}}$ can be obtained by localized collision-knot-recognition events; however, the mechanism of Ref. 8 is unable to explain values of P_{knot} as small as those observed experimentally.

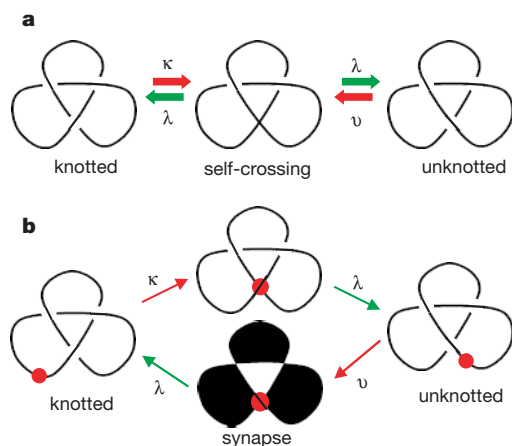


Figure 1 Simplest kinetic models of type II topoisomerases. **a**, Loop of DNA capable of freely crossing itself (a 'ghost' or 'phantom' polymer, such as a strand of DNA that can intermittently break and rejoin^{6,7}) spends some fraction of time $P_{\text{knot}}^{\text{eq}}$ as a knotted strand and some fraction $P_{\text{unknot}}^{\text{eq}}$ as an unknotted strand. For the 3–10-kb DNA strands used here, knot–unknot free energy differences are entropic, and the strands at the self-crossing point (the X intersection) are not subject to forces that determine knotting or unknotting. Therefore, transitions away from the self-crossing point occur at equal rates λ for knotted and unknotted strands (green arrows); any knot–unknot discrimination must be based on the self-crossing rates κ and ν (red arrows). **b**, A more realistic 'one-way' or 'two-gate'¹² strand passage model of a type II topoisomerase distinguishes two DNA–topoisomerase–DNA 'synapse' states. On isolated circular DNA strands, transitions occur from knotted to unknotted states at a rate $\kappa\lambda$, and back again at a rate $\nu\lambda$. For a topoisomerase that does not consume stored energy, $(\nu\lambda)/(\kappa\lambda)$ must equal the thermal equilibrium ratio of knotted to unknotted strands for 'ghost' DNA as in **a**. If the topoisomerase uses stored energy, it is possible for the knot probability to be reduced below that expected at thermal equilibrium.

We propose that the type II topoisomerases described above suppress entanglements using a type of 'kinetic proofreading'. This was first discussed in general terms by Hopfield² and Ninio³, who showed how the release of energy (for example, from ATP hydrolysis) could be used to make molecular recognition processes more specific than one would expect at thermal equilibrium. Figure 4 shows our proofreading scheme for type II topoisomerases interacting with DNA plasmids which may be knotted or unknotted. For simplicity, only one type of knot is considered; for the DNA strands of ref. 1, essentially all the knots are expected to be trefoils⁹. We limit our discussion to DNA-disentangling topoisomerases (such as eukaryote type II topoisomerases and *E. coli* type IV topoisomerases) that pass distant, disjointed DNA segments through one another.

We now describe our model in the context of the relatively simple knotting–unknotting single-DNA reaction. Linking–unlinking of two circular DNA strands can be discussed in a similar way for the dilute solution conditions of ref. 1. Starting from either knotted or unknotted strands with a topoisomerase attached (Fig. 4, state 1), a synapse can reversibly occur (Fig. 4, $1 \leftrightarrow 2$; numbers indicate how many DNA segments are bound by the topoisomerase). For ~10-kb plasmids we may assume that the rate of synapse formation on knotted strands exceeds the formation rate on unknotted strands by an amount of at least $P_{\text{unknot}}^{\text{eq}}/P_{\text{knot}}^{\text{eq}}$, given that it is possible for a single local recognition step to be used to obtain $P_{\text{knot}} < P_{\text{knot}}^{\text{eq}}$ (ref. 8). As the molecular motions leading to synapsis are the only dependents of these transitions on the knotting state (that is, the off-rate λ values are likely to be determined by the energetics of the topoisomerase–DNA binding rather than by global DNA conformation), we assume that the synapsis rates for knots (κ) and unknots (ν) satisfy $\nu/\kappa \leq P_{\text{knot}}^{\text{eq}}/P_{\text{unknot}}^{\text{eq}}$.

From state 2 (Fig. 4), an irreversible transition can occur to a transient state 1* with only one of the DNA segments bound to an 'activated' topoisomerase (the other DNA segment is released). Irreversibility could be enforced with ATP binding, possibly coupled to topoisomerase conformation change and cleavage of the bound DNA. As no overall DNA conformational change is involved in this step, the rate α should be insensitive to DNA topology.

The state 1* can 'decay' back to 1 at a topology-independent decay rate γ , giving spontaneous 'deactivation' of the topoisomerase with no change in DNA topology. Alternatively, a new synapse can reversibly form between a new DNA segment and the activated topoisomerase–DNA complex ($1^* \leftrightarrow 2^*$), which will lead to strand

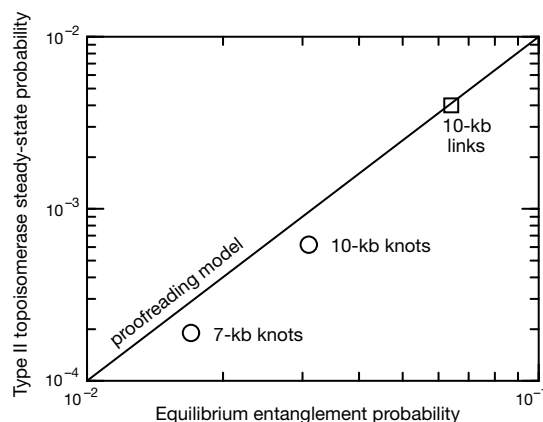


Figure 2 Experimental knotting (circles) and linking (square) probabilities from ref. 1, compared with our model. The horizontal axis shows the thermal equilibrium entanglement probability; the vertical axis shows the steady-state entanglement probability results in the presence of type II topoisomerases and ATP. The kinetic proofreading model (see text) is able to reduce the knotting probability to levels below the solid line, which is defined by (steady-state) = (equilibrium)².

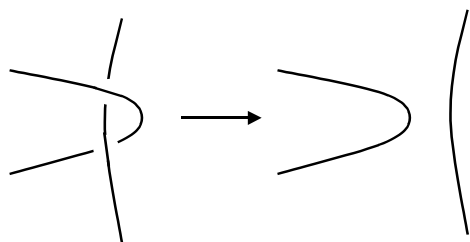


Figure 3 Nonequilibrium bend-recognizing topoisomerase mechanism of Vologodskii. Juxtapositions of sharp DNA bends with straight segments (left) occur more frequently on knotted DNA strands than on unknotted DNA strands⁸. By allowing only the forward strand passage (arrow) to occur, the knotting probability can be forced below that expected at thermal equilibrium. Therefore, knotted strands can be 'recognized' by topoisomerases on the basis of localized DNA geometry.

transfer. As before, the off-rate λ' values are assumed to be insensitive to DNA knotting. As we are only concerned with topology-changing events, we may assume that the ratio of the recollision rates for knots (κ') and unknots (v') is $v'/\kappa' = P_{\text{knot}}^{\text{eq}}/P_{\text{unknot}}^{\text{eq}}$. We note that if the second synapse again requires specific types of collisions, such as that of Fig. 3, we could have $v'/\kappa' < P_{\text{knot}}^{\text{eq}}/P_{\text{unknot}}^{\text{eq}}$.

Finally, irreversible transitions from $2^* \rightarrow 1$ achieve strand passage, DNA religation, and release of the passed segment. At the end of the $2^* \rightarrow 1$ transitions, the topoisomerase is reset and ready for another cycle. ATP hydrolysis and product release could be involved in this step. These irreversible transitions should be controlled by local topoisomerase–DNA interactions, and should occur at a strand passage rate μ that is independent of DNA topology.

The steady-state knot–unknot ratio for this model is:

$$\frac{P_{\text{knot}}}{P_{\text{unknot}}} = \frac{(\gamma[\lambda' + \mu] + \kappa'v)v'}{(\gamma[\lambda' + \mu] + v'\mu)\kappa\kappa'} \quad (1)$$

We assume that $\gamma[\lambda' + \mu] \gg \kappa'\mu$, as the recollision rate κ' requires a particular outcome of large-scale DNA conformational change, whereas the strand passage rate μ includes the process of ATP hydrolysis product release, both of which are expected to be slow relative to the rate of DNA release λ' , and the decay rate γ . Similarly, $\gamma[\lambda' + \mu] \gg v'\mu$. These conditions simplify equation (1) to:

$$\frac{P_{\text{knot}}}{P_{\text{unknot}}} = \frac{v'}{\kappa\kappa'} \leq \left(\frac{P_{\text{knot}}^{\text{eq}}}{P_{\text{unknot}}^{\text{eq}}} \right)^2 \quad (2)$$

This shows that the reaction of Fig. 4 can reduce the knotting probability below the square of $P_{\text{knot}}^{\text{eq}}/P_{\text{unknot}}^{\text{eq}}$, as is observed experimentally (Fig. 2). We note that all of the topology-independent rate constants have dropped out of the final knot probability, leaving a result with no adjustable parameters. The topology-dependent rates v , v' , κ and κ' could be roughly estimated using molecular dynamics simulations.

The product of v/κ and v'/κ' appears in equation (1) because the inputs to the second synapse are biased by the outcome of the first synapse; the second stage 'proofreads' the first. For $P_{\text{knot}}^{\text{eq}} \ll P_{\text{unknot}}^{\text{eq}}$, as in experiments on 10-kb and 7-kb plasmids, our model explains how knots are so effectively removed by topoisomerases. Type II topoisomerases are also capable of suppressing the linkage probability to roughly the square of that expected for thermal equilibrium (Fig. 2). Further, the distribution of supercoiling in the presence of type II topoisomerases corresponds roughly to a squaring of the equilibrium distribution. This suggests that the discrimination step ($1 \leftrightarrow 2 \rightarrow 1^*$ in our model) recognizes knots, catenanes and supercoils; simulations⁸ and structural studies of topoisomerases are needed to understand this in detail.

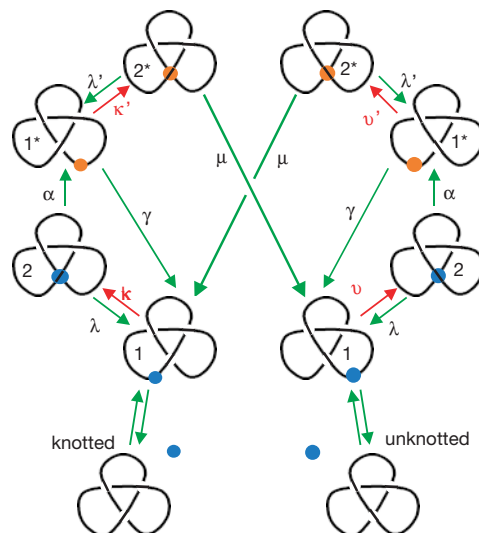


Figure 4 Proposed kinetic model for type II topoisomerase using kinetic proofreading of DNA topology. We note that reactions of the form of Fig. 1b occur twice along the knotting and unknotting pathways. The number of DNA segments bound to the topoisomerase is shown for each loop. 'Activated' topoisomerases (indicated with an asterisk) are able to pass DNA through DNA (see text). The topoisomerase itself is shown in blue when inactive and red when active. By cascading two synapsis events separated by irreversible transitions, the first synapsis ($1 \leftrightarrow 2$) delivers an excess of knotted strands over unknotted strands to the second synapsis ($1^* \leftrightarrow 2^*$); the second reaction 'proofreads' the first. Proofreading can reduce the knot–unknot ratio to below $(P_{\text{knot}}^{\text{eq}}/P_{\text{unknot}}^{\text{eq}})^2$. Most of the transitions do not depend on knottedness (green arrows); all discrimination of topology is based on synapse formation rates (red arrows), as in Fig. 1. The topology-independent rate constants do not contribute to the final knot–unknot steady-state ratio of this model.

The model of Fig. 4 is compatible with experimental constraints. The strand passage pathway is 'one-way', in accord with the experimentally supported¹⁰ 'two-gate' model for type II topoisomerases in which the passed strand enters and exits the DNA–topoisomerase complex through different gateways. Other experiments show that a single round of strand passage can occur when non-hydrolysable ATP analogues are used^{11,12}, and support a ATP-binding-driven DNA-clamp model of topoisomerase function¹⁰.

Our proofreading reaction is compatible with the DNA-clamping model. For example, the $2 \rightarrow 1^*$ transition could correspond to ATP binding, stimulated by the first synapsis. The decay of the 1^* state (γ) could then correspond to the closing of the DNA clamp. The second synapsis would then correspond to DNA recollisions occurring before this topoisomerase conformational change. Non-hydrolysable ATP would block completion of the $2^* \rightarrow 1$ transition, trapping the topoisomerase in an inactive state after one strand passage. Alternatively, given recent experiments indicating two sequential ATP hydrolysis events^{13,14}, it is tempting to imagine that the first ATP hydrolysis is somehow involved in the $2 \rightarrow 1^*$ transition; however, this is difficult to reconcile with clamp closure triggered by ATP binding.

Finally, proofreading may increase knotting under conditions where knotted strands are more likely than unknotted strands at thermal equilibrium, where topology-changing self-collisions on unknotted strands occur at higher frequency than on knotted strands. This is the case for large DNA strands: ~200-kb plasmids with equilibrated topology have more trefoils than any other type of knots, including unknotted strands⁹, and on these molecules type II topoisomerases should generate more trefoils and fewer unknotted strands than expected in equilibrium. Avoidance of a situation in which topoisomerase becomes knot-generating rather than knot-removing suggests arranging large DNA strands into multiple 'loop' domains of less than 100 kb to ensure entanglement removal, as

'links' between the different loop domains. This provides a rationale for organization of, for example, the 4.5-Mb chromosome of *E. coli* into loop domains of approximately this size^{15,16}. □

Received 22 December 1998; accepted 18 August 1999.

- Rybenkov, V. V., Ullsperger, C., Vologodskii, A. V. & Cozzarelli, N. R. Simplification of DNA topology below equilibrium values by type II topoisomerases. *Science* **277**, 690–693 (1997).
- Hopfield, J. J. Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proc. Natl Acad. Sci. USA* **71**, 4135–4139 (1974).
- Ninio, J. Kinetic amplification of enzyme discrimination. *Biochimie* **57**, 587–595 (1975).
- Frank-Kamenetskii, M. D., Lukashin, A. V. & Vologodskii, A. V. Statistical mechanics and topology of polymer chains. *Nature* **258**, 398–402 (1975).
- Klenin, K. V., Vologodskii, A. V., Anshelevich, V. V., Dykhne, A. M. & Frank-Kamenetskii, M. D. Effect of excluded volume on topological properties of circular DNA. *J. Biomol. Struct. Dyn.* **5**, 1173–1185 (1988).
- Shaw, S. Y. & Wang, J. C. Knotting of a DNA chain during ring closure. *Science* **260**, 533–536 (1993).
- Rybenkov, A. V., Cozzarelli, N. R. & Vologodskii, A. V. Probability of DNA knotting and the effective diameter of the DNA double helix. *Proc. Natl Acad. Sci. USA* **90**, 5307–5311 (1993).
- Vologodskii, A. V. in RECOMB 98: Proceedings of the second annual international conference on computational molecular biology 266–269 (Association for Computing Machinery, New York, 1998).
- Deguchi, T. & Tsurusaki, K. A statistical study of random knotting using the Vassiliev invariants. *J. Knot Theory Ramif.* **3**, 321–353 (1994).
- Roca, J. & Wang, J. C. DNA transport by a type II DNA topoisomerase: evidence in favor of a two-gate mechanism. *Cell* **77**, 609–616 (1994).
- Roca, J. & Wang, J. C. The capture of a DNA double helix by an ATP-dependent protein clamp: a key step in DNA transport by type II DNA topoisomerases. *Cell* **71**, 833–840 (1992).
- Roca, J., Berger, J. M. & Wang, J. C. On the simultaneous binding of eukaryotic DNA topoisomerase II to a pair of double-stranded DNA helices. *J. Biol. Chem.* **268**, 14250–14255 (1993).
- Harkins, T. T. & Lindsley, J. E. Pre-steady-state analysis of ATP hydrolysis by *Saccharomyces cerevisiae* DNA topoisomerase II. 1. A DNA-dependent burst in ATP hydrolysis. *Biochemistry* **37**, 7292–7299 (1998).
- Harkins, T. T. & Lindsley, J. E. Pre-steady-state analysis of ATP hydrolysis by *Saccharomyces cerevisiae* DNA topoisomerase II. 2. Kinetic mechanism for the sequential hydrolysis of two ATP. *Biochemistry* **37**, 7299–7312 (1998).
- Worcel, A. & Burgi, E. On the structure of the folded chromosome of *Escherichia coli*. *J. Mol. Biol.* **71**, 127–147 (1972).
- Sinden, R. R. & Pettijohn, D. E. Chromosomes in living *Escherichia coli* cells are segregated into domains of supercoiling. *Proc. Natl Acad. Sci. USA* **78**, 224–228 (1981).

Acknowledgements

We thank D. Chatenay, N. R. Cozzarelli, G. B. Mindlin, V. Rybenkov, E. D. Siggia, A. V. Vologodskii and E. L. Zechiedrich for discussions. J.F.M. and J.Y. acknowledge the support of the NSF, the Research Corporation, the Petroleum Research Fund, and the Whitaker Foundation. M.O.M. acknowledges support of the Sloan Foundation and the Mathers Foundation.

Correspondence and requests for materials should be addressed to J.F.M. (e-mail: jmarko@uic.edu).

A triple β -spiral in the adenovirus fibre shaft reveals a new structural motif for a fibrous protein

Mark J. van Raaij*, Anna Mittraki†, Gilles Lavigne† & Stephen Cusack*

* European Molecular Biology Laboratory, Grenoble Outstation, c/o Institut Laue Langevin, BP 156, 38042 Grenoble Cedex 9, France

† Institut de Biologie Structurale (CEA-CNRS), 41 rue Jules Horowitz, 38027 Grenoble Cedex 1, France

Human adenoviruses¹ are responsible for respiratory, gastroenteric and ocular infections² and can serve as gene therapy vectors³. They form icosahedral particles with 240 copies of the trimeric hexon protein arranged on the planes and a penton complex at each of the twelve vertices. The penton consists of a pentameric base, implicated in virus internalization⁴, and a protruding trimeric fibre, responsible for receptor attachment⁵. The fibres are homo-trimeric proteins containing an amino-terminal penton base attachment domain, a long, thin central shaft and a carboxy-terminal cell attachment or head domain. The

Table 1 Crystallographic data and refinement statistics

Data and refinement statistics (values in parentheses are for the resolution bin 2.53–2.40 Å)	
Space group	C2
Cell dimensions	$a = 165.51 \text{ Å}$, $b = 95.87 \text{ Å}$, $c = 211.77 \text{ Å}$, $\beta = 106.83^\circ$ 25–2.4 Å
Resolution range	103,327 (10,696)
No. of reflections	84.3% (59.9%)
Completeness	2.7 (1.7)
Multiplicity	0.132 (0.287)
R_{merge}	0.232 (0.312)
R -factor	0.265 (0.365)
R_{free} value	12,623
No. of atoms	101,663
No. of reflections used in refinement	1,662
No. of reflections used for R_{free}	73.1%
Solvent content	12,042
No. of protein atoms (6×264 residues)	581
No. of water molecules	35.3 Å ²
B -value from Wilson plot (3.5–2.4 Å)	35.2 Å ²
Mean B -value	35.0 Å ²
Average protein B -value	38.0 Å ²
Average water B -value	
Ramachandran plot of non-glycine and non-proline residues	
Most favourable regions	1,102 (78.5%)
Additional allowed regions	275 (19.6%)
Generously allowed regions	17 (1.2%)
Disallowed regions	10 (0.7%)
r.m.s. deviations from ideal values	
Bond distances	0.006 Å
Angles	1.4°
$R_{\text{merge}} = \sum_i \sum_j I_{ijk} - \langle I_{ijk} \rangle / \sum_i \sum_j \langle I_{ijk} \rangle$ where the sum i is over all separate measurements of the unique reflections hkl .	
R -factor = $\sum_{hkl} F_{\text{obs}} - F_{\text{calc}} / \sum_{hkl} F_{\text{obs}} $	
R_{free} , as R -factor but summed only over the test reflections	
B -value = $8\pi^2 \langle u^2 \rangle$ where $\langle u^2 \rangle$ is the apparent mean square deviation from the atomic position.	

shaft domain contains a repeating sequence motif with an invariant glycine or proline and a conserved pattern of hydrophobic residues⁶. Here we describe the crystal structure at 2.4 Å resolution of a recombinant protein containing the four distal repeats of the adenovirus type 2 fibre shaft plus the receptor-binding head domain. The structure reveals a novel triple β -spiral fibrous fold for the shaft. Implications for folding of fibrous proteins (misfolding of shaft peptides leads to amyloid-like fibrils) and for the design of a new class of artificial, silk-like fibrous materials are discussed.

The human adenovirus serotype 2 (Ad2) fibre is a trimer of 582 residues per monomer⁷, of which the head domain is essential for trimerization and autonomously trimerizes when expressed⁸. The high-resolution structures of the heads of Ad2 and Ad5 are known^{9,10} and are very similar: each head monomer forms an eight-stranded anti-parallel β -sandwich structure and the three monomers interact to form a three-bladed propeller. The head domain is responsible for binding to the cell receptor, which has been identified to be a human protein of unknown function: the coxsackievirus and adenovirus receptor¹¹. This protein serves as the receptor for coxsackieviruses of subgroup B and adenoviruses of all subgroups except subgroup B (ref. 12). The primary sequence of the fibre shaft consists of 15-residue pseudo-repeats (22 of them for Ad2 and Ad5, ref. 7). Green *et al.*⁶ predicted that these repeats contain two β -strands and two turns (the cross- β model). Stouten *et al.*¹³ subsequently proposed a triple β -helical model, taking into account length measurements from electron microscopy and fibre diffraction patterns¹⁴. The full-length fibre is very stable, resistant to heat (its melting temperature is 85 °C, ref. 15) and detergents (at low temperatures¹⁶), and the shaft domain is highly resistant to proteases¹⁷.

The Ad2 fibre unfolds through a stable intermediate in which the C-terminal head and distal part of the shaft remain folded and trimeric¹⁶. The stable domain has been identified to span residues 319–582 and has been cloned and expressed in *Escherichia coli* (M.J.v.R. *et al.*, unpublished results). The recombinant protein was

The shaft structure reveals a new fold, which we call a ‘triple β -spiral’ (Fig. 1). The repeating motif of 15 residues (a–o) exists in two types, with either a proline or a glycine at position j (Fig. 2). The four repeats in our structure are all of the glycine type. Each repeat comprises an extended strand (residues b–h) which runs parallel to the fibre axis, followed by a type 2 β -turn, containing the conserved glycine with ϕ around 80° and ψ around 0° (ϕ and ψ are polypeptide chain backbone dihedral angles). A proline at this position may impose a type 1 turn (ϕ around -90° ; and ψ around 0°). The turn is followed by another β -strand (residues k–n) which runs backwards at about 45° to the shaft axis. The repeats are joined together by a solvent-exposed variable loop (residues o–a). The intra-chain transformation from one repeat to the next is a displacement along the shaft axis of about $13 \text{ \AA}$ and clockwise rotation of just over 50° (as seen from the N to the C terminus); the axial displacement agrees with previous measurements by electron microscopy¹⁹. Three conserved hydrophobic residues in each repeat (residues c, e and k in Fig. 2; shown in orange in Fig. 3a) point towards the central three-fold axis, forming a longitudinal hydrophobic core with the equivalent

The resulting structure is a complex but regular, highly cross-linked structure which, together with a high proportion of buried surface (one-third of solvent-accessible surface of a shaft monomer is buried upon trimer formation), accounts for the high rigidity and stability of the shaft. The average diameter of the shaft, excluding the surface loops, is 15 Å, and the radius of the large loop (residues 346–355) is 22 Å. Figure 4 shows a model of a fibre 300 Å long (the full length of the Ad2 fibre) to give an impression of the proportions of the molecule. The fold of the adenovirus shaft is unique amongst protein structures so far determined and different from both the cross- β model⁶ and the triple-helical model¹³. The difference lies largely in the fact that the β -strands are oriented more along the fibre axis than transverse to it, as previously assumed. Further, the intra-chain hydrogen bonding pattern shown in Fig. 3b suggests that the basic repeating structural module should be redefined as starting from residue g and ending at residue f. This redefinition naturally leads to a clear end of the regular shaft structure at residue Gly 392. This residue is followed by a linker region between the shaft and the head domains (residues 393–398), which was originally defined as part of repeat number 22 of the shaft. The electron density is not well defined for residues 394–396, suggesting that the linker is flexible. There is also a solvent-filled volume around the three-fold axis between the head and shaft, rather than a continuation of the longitudinal hydrophobic core (Fig. 1a). In the crystal,

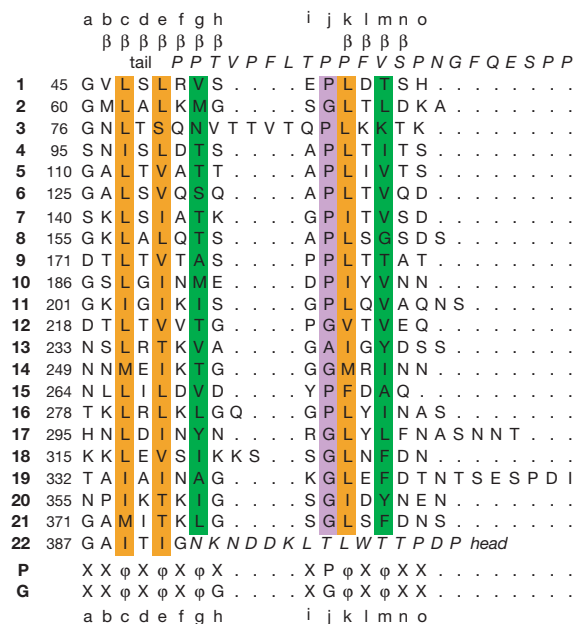


Figure 2 Alignment of the 22 repeats¹⁵ in the Ad2 shaft. The repeating residues are labelled a to o. Residues forming β -strands are indicated with β . The conserved glycines or prolines in the tight turn (purple, except the corresponding residue in repeat 13, which is an alanine), residues involved in the hydrophobic core (orange), and residues forming the peripheral hydrophobic patches (green) are shown. Some residues of the fibre tail domain (23–44) are at the top (*italics*). Repeat 3 (shown in *italics*) cannot be aligned reliably and may have a different structure. The spacer (393–NKNDK–398) between the shaft and the head and the first nine residues of the head domain are also in *italics*. The consensus sequences of the proline and glycine repeats (bottom, P and G) are shown (ϕ is hydrophobic).

the three-fold shaft axis is tilted with respect to the three-fold axis running through the head domain by $\sim 2^\circ$, further confirming this flexibility. This implies the possibility of relative movement between the head and shaft, which may be required to optimise receptor binding. Interestingly, this spacer region is not conserved in adenoviruses of subgroup B, which use a different, as yet unknown, primary receptor¹².

Comparisons of adenovirus serotypes⁷ show that the shaft sequence can accommodate sequence variations, insertions and deletions. The most variable element is the solvent-exposed loop, for which this structure shows examples of 4, 5 and 12-residue turns. This loop can contain insertions of up to eight residues and could possibly be used to insert foreign sequences (potentially useful for gene therapy), while preserving a correctly folded fibre. Whereas other fibrous proteins contain repetitive sequences, the presence of a variable loop appears to be a unique feature of the adenovirus fibre shaft fold. In the Ad2 sequence, the longer turns are concentrated at the C-terminal end of the shaft and may constitute serotype-specific antigenic regions, as the sequence of these loops varies between Ad2 and Ad5. The compact glycine or proline loops are more highly conserved four-residue turns, with some exceptions, such as the insertion of two residues (323-Lys-Ser-324) in the glycine turn of repeat 18 in this structure. Amplification of insertions in this region occurs in some non-human adenovirus shaft sequences^{20,21}. The breakdown of the consensus sequence in the third repeat (Fig. 2) has been correlated to a kink in the shaft at this point, as observed by electron microscopy¹⁹. Similar putative hinge regions correlated with aberrant repeats have been described in the very long bovine adenovirus fibre shaft²⁰ and in certain avian adenovirus fibres²¹. High stability, length and rigidity, together with well-defined hinge positions in the shaft or at the head-shaft

boundary, are clearly desirable properties for an extra-cellular viral protein required to bind to particular receptors amongst many other cell surface molecules.

Most known fibrous protein structures are based on α -helical coiled coils or the collagen triple helix²². Some elongated proteins such as the bacteriophage P22 tailspike²³ and pectate lyase²⁴ are based on a β -helix. The triple β -spiral is quite distinct from these. It is also very different from the crystalline structural region in silk²⁵. However, sequence similarities shared with T4 bacteriophage short and long fibres indicate that this novel structural motif may exist in other proteins. The protein sequences of the bacteriophage T4 short and long fibre shafts show consensus repeats of 17 residues (xTxxxxG ϕ ϕ X ϕ xTxxx ϕ and xGxxxx ϕ ϕ TPxx ϕ xxx; x is any amino acid, ϕ is hydrophobic), which contain two predicted β -strands connected by a predicted β -turn in which a glycine or proline is conserved²⁶. We note that the positions of the conserved hydrophobic residues relative to the conserved glycine or proline are not the same as in the adenovirus shaft motif, so some structural differences must also exist. It remains to be seen how many members there are of this new structural class of fibrous proteins.

To assemble into the triple β -spiral, the shaft repeats need to be brought into the correct registration. Synthetic peptides containing shaft residues 355–395 fail to assemble correctly, but aggregate,

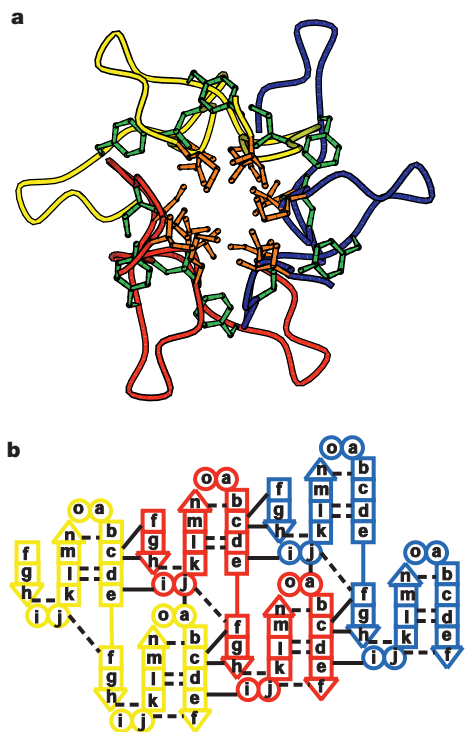


Figure 3 Interactions in the shaft. **a**, Hydrophobic core. The main chain of residues 356–389 of monomers A (red), B (blue) and C (yellow) is shown. Residues contributing to the central hydrophobic core (orange) and residues contributing to the peripheral hydrophobic patches (green) are also given. **b**, Diagram of the hydrogen-bonding network within one monomer and between adjacent monomers. Residues are labelled according to Fig. 2. Intra-chain (dotted lines) and inter-chain hydrogen bonds (solid black lines) that are conserved in each of the four repeats of the structure are shown.

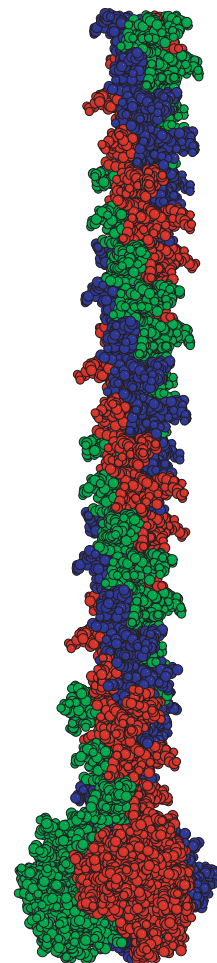


Figure 4 Space-filled representation of a model containing 21 repeats of the shaft. The model contains residues 321–582 of chains A (red), B (blue) and C (green), to which copies of residues 361–376 (one repeat) were added by repeated application of the inter-motif transformation to obtain a shaft length of 21 repeats or 265 Å (about the length of the entire Ad2 fibre shaft). We note that seven inter-motif transformations provide a rotation through 360° .

probably through out-of-register interactions, and form fibrils similar to amyloid fibrils (A.M. *et al.*, unpublished results). The head domain probably acts as the registration signal for correct folding, as do the C-terminal domains of T4 fibrin²⁷ and pro-collagen peptides. Our structural information, combined with folding studies, could be used to clarify the mechanisms that control correct assembly or misassembly in fibrous β -sheet proteins, as was done for the P22 tailspike²⁸. Because of its stability and morphology, the adenovirus fibre has served as a model for synthetic fibre design. O'Brien *et al.*²⁹ have designed recombinant polymers using simplified consensus adenovirus shaft repeats as building blocks, which were bacterially expressed as inclusion bodies, refolded, purified and spun into fibres. In one case, using the sequence NAL-RIKLGSGLDNFN as a 26-fold repeating motif, they obtained ordered fibres with mechanical properties comparable to several commercial textile fibres. Our structural information on the fibre shaft may stimulate further design of synthetic fibres with improved or novel properties, for example by engineering a simple trimerization domain coupled to a certain number of shaft repeats to obtain high tensile strength fibres of a defined length. □

Methods

Crystallization and structure solution

A C-terminal part of the adenovirus type 2 fibre protein containing residues 319–582 was expressed, purified and crystallized (M.J.v.R. *et al.*, unpublished data). Crystals were obtained using the sitting-drop vapour-diffusion method by combining 1 μ l of protein solution (14 mg ml⁻¹) and 1 μ l precipitant (0.8 M ammonium sulphate, 0.1 M sodium citrate, pH 4.0) and equilibrating at 20 °C against the precipitant solution. A cluster of plate-like crystals was obtained after six months; a single crystal (0.4 × 0.15 × 0.05 mm) was obtained by transferring the cluster to the precipitant solution containing 25% by volume glycerol and separating it with a glass hair probe. It was flash-frozen in a cryo-stream at 100 K. Data was collected on station ID14eh3 at the European Synchrotron Radiation Facility. Diffraction images were processed using CCP4 programs³⁰. The structure was solved by molecular replacement (AMORE³⁰) using the Ad2 fibre head structure⁸. Apart from a side-chain flip of Phe 480, the structure of the head domain does not change in the presence of the shaft, although it is now apparent that residues 388–395, which form the putative last repeat of the shaft, were forced into an unnatural conformation by crystal packing in the structure of the head alone. For calculation of the free R-factor, a subset of reflections was selected in thin shells of resolution. No non-crystallographic symmetry restraints were used in the refinement as crystal contacts result in the six monomer chains being slightly different.

Received 18 June; accepted 26 August 1999.

1. Schenk, T. in *Fields Virology* 2111–2148 (Lippincott-Raven, Philadelphia, 1996).
2. Horwitz, M. S. in *Fields Virology* 2149–2171 (Lippincott-Raven, Philadelphia, 1996).
3. Robbins, P. D., Tahara, H. & Ghivizzani, S. C. Viral vectors for gene therapy. *Trends Biotechnol.* **16**, 35–40 (1998).
4. Wickham, T. J., Mathias, P., Cheres, D. A. & Nemerow, G. R. Integrins $\alpha\beta 3$ and $\alpha\beta 5$ promote adenovirus internalization but not virus attachment. *Cell* **73**, 309–319 (1993).
5. Philipson, L., Lonberg-Holm, K. & Pettersson, U. Virus-receptor interaction in an adenovirus system. *J. Virol.* **2**, 1064–1075 (1986).
6. Green, N.M., Wrigley, N. G., Russel, W. C., Martin, S. R. & McLachlan, A. Evidence for a repeating cross β structure in the adenovirus fibre. *EMBO J.* **2**, 1357–1365 (1983).

7. Chroboczek, J., Ruigrok, R. W. H. & Cusack, S. Adenovirus fiber. *Curr. Top. Microbiol. Immunol.* **199**, 163–200 (1995).
8. Hong, J. S. & Engler, J. A. Domains required for assembly of adenovirus type 2 fiber trimers. *J. Virol.* **70**, 7071–7078 (1996).
9. van Raaij, M. J., Louis, N., Chroboczek, J. & Cusack, S. Structure of the human adenovirus serotype 2 fiber head domain at 1.5 Å resolution. *Virology* **262**, 333–343 (1999).
10. Xia, D., Henry, L. J., Gerard, R. D. & Deisenhofer, J. Crystal structure of the receptor-binding domain of adenovirus type 5 fiber protein at 1.7 Å resolution. *Structure* **2**, 1259–1270 (1994).
11. Bergelson, J. M. *et al.* Isolation of a common receptor for coxsackie B viruses and adenoviruses 2 and 5. *Science* **275**, 1320–1323 (1997).
12. Roelvink, P. W. *et al.* The coxsackievirus-adenovirus receptor protein can function as a cellular attachment protein for adenovirus serotypes from subgroups A, C, D, E, and F. *J. Virol.* **72**, 7909–7915 (1998).
13. Stouten, P. F. W., Sander, C., Ruigrok, R. W. H. & Cusack, S. New triple-helical model for the shaft of the adenovirus fibre. *J. Mol. Biol.* **226**, 1073–1084 (1992).
14. Devaux, C., Adrian, M., Berthet-Colominas, C., Cusack, S. & Jacrot, B. Structure of adenovirus fibre. I. Analysis of crystals of fibre from adenovirus serotypes 2 and 5 by electron microscopy and X-ray crystallography. *J. Mol. Biol.* **215**, 567–588 (1990).
15. Louis, N. *Etude de la structure de la fibre de l'adénovirus humain de la sérotype 2; son devenir et son rôle au cours de l'infection*. PhD thesis, Univ. Joseph-Fourier, Grenoble (1994).
16. Mitaki, A. *et al.* Unfolding studies of human adenovirus type 2 fiber trimers: evidence for a stable domain. *Eur. J. Biochem.* **264**, 599–606 (1999).
17. Devaux, C., Caillat-Boudin, M. L., Jacrot, B. & Boulanger, P. Crystallization, enzymatic cleavage, and the polarity of the adenovirus type 2 fiber. *Virology* **161**, 121–128 (1987).
18. Brunger, A. T. *et al.* Crystallography & NMR System: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
19. Ruigrok, R. W. H., Barge, A., Albiges-Rizo, C. & Dayan, S. Structure of adenovirus fibre. II. Morphology of single fibres. *J. Mol. Biol.* **215**, 589–596 (1990).
20. Ruigrok, R. W. H., Barge, A., Mittel, S. K. & Jacrot, B. The fibre of bovine adenovirus type 3 is very long but bent. *J. Gen. Virol.* **75**, 2069–2073 (1994).
21. Hess, M., Cuzange, A., Ruigrok, R. W. H., Chroboczek, J. & Jacrot, B. The avian adenovirus penton: two fibres and one base. *J. Mol. Biol.* **252**, 379–385 (1995).
22. Beck, B. & Brodsky, B. Supercoiled protein motifs: The collagen triple-helix and the α -helical coiled coil. *J. Struct. Biol.* **122**, 17–29 (1998).
23. Steinbacher, S. *et al.* Crystal structure of P22 tailspike protein: interdigitated subunits in a thermostable trimer. *Science* **265**, 383–386 (1994).
24. Yoder, M. D., Keen, N. T. & Jurnak, F. New domain motif: Structure of pectate lyase C, a secreted plant virulence factor. *Science* **260**, 1503–1507 (1993).
25. Marsh, R. E., Corey, R. B. & Pauling, L. The structure of Tussah silk fibroin. *Acta Crystallogr.* **8**, 710–715 (1955).
26. Cerritelli, M. E., Walls, J. S., Simon, M. N., Conway, J. F. & Steven, A. C. Stoichiometry and dominant organization of the long-fiber of bacteriophage T4: A hinged viral adhesion. *J. Mol. Biol.* **260**, 767–780 (1996).
27. Tao, Y., Strelkov, S. V., Mesyanshinov, V. & Rossmann, M. G. Structure of bacteriophage T4 fibrin: A segmented coiled coil and the role of the C-terminal domain. *Structure* **5**, 789–798 (1997).
28. Betts, S. & King, J. There's a right and a wrong way: in vivo and in vitro folding, misfolding and subunit assembly of the P22 tailspike. *Structure* **7**, R131–R139 (1999).
29. O'Brien, J. P. *et al.* in *Silk Polymers, Materials Science and Biotechnology* 104–117 (ACS Symposium Series 544, American Chemical Society, Charlottesville, Virginia, 1993).
30. Collaborative Computational Project Number 4. The CCP4 suite: Programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).

Acknowledgements

We thank P. Goeltz and N. Cohet for help with cloning, expression and purification and J. Gagnon for comments. M.J.v.R. was supported by an EU Biotech II fellowship. The EMBL-ESRF Joint Structural Biology Group provided synchrotron radiation facilities.

Correspondence and requests for materials should be addressed to S.C. (e-mail: cusack@embl-grenoble.fr). Atomic co-ordinates and structure factors are available from the Protein Data Bank under accession numbers 1QIU and 1QIUSE, respectively.